SOCIAL NETWORKS AND IMMIGRATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dafeng Xu

January 2017

SOCIAL NETWORKS AND IMMIGRATION

Dafeng Xu, Ph.D.

Cornell University 2017

My dissertation focuses on the intersection between immigration and social networks. The three essays of the dissertation study network formation, network characteristics, and network effects, respectively. The first essay investigates how immigrants construct carpooling networks in order to deal with language problems when commuting. I focus on the role of language proficiency, and find that immigrants with lower levels of English skills are more likely to commute to work by carpooling. Similarly, the number of co-riders is negatively associated with English proficiency. In other words, immigrants create need-based carpooling networks in order to tackle potential language problems. The second essay studies how social networks can be defined based on a typical acculturational behavior, namely, English-name usage. Exploiting a natural linguistic experiment, I find that Chinese students with English-name usage have more close friends who are also English-name users. This implies that homophily could occur among friendships within the same ethnic group in the context of immigration. The third essay examines social network effects among highly professional migrants: I focus on French football players in England and study whether ethnic networks affect yearly migration outcomes. I find that a player exposed to a larger French network is more likely to stay in England, although not necessarily the same team. However, the network effects are highly heterogeneous, and ethnic networks do not always benefit those who need support most, such as veteran players or players with relatively low levels of outputs.

BIOGRAPHICAL SKETCH

The author of this dissertation, Dafeng Xu, is a PhD candidate in the Department of City and Regional Planning, Cornell University. He is also affiliate with Cornell Population Center and will earn the PhD minor in demography. Prior to coming to Cornell, Dafeng Xu received his master's degree in systems engineering from the University of Pennsylvania in 2013, and his bachelor's degree in computer science from Peking University in 2011. He will join the University of Minnesota as the postdoctoral researcher after graduating from Cornell.

To Zhuzhu the Arctic,

my beloved piglet.

ACKNOWLEDGEMENTS

I greatly acknowledge my committee chair, Nancy Brooks, as well as Lawrence Kahn, Ravi Kanbur, and Daniel Lichter who kindly served on my committee, for their support and encouragement. They gave me consistently good advice on my dissertation and other research projects, as well as tremendous freedom to explore my academic interests in many topics and fields. Three years ago, I arrived at Cornell and started my doctoral studies as an ambitious yet inexperienced researcher with almost no knowledge of geography, economics, and demography, and they have really taught me a lot. I will be forever grateful for their guidance. Special thanks go to my master's advisor at Penn, Tony Smith, who led me to this discipline and has always been my reader. None of this would have been possible if it were not for his continued support.

Cornell is a huge interdisciplinary academic community and I am extremely fortunate to have conversations and discussions with many Cornell professors who were not on my committee. I would especially like to thank Christopher Anderson, Francine Blau, Lawrence Blume, Kieran Donaghy, Maria Fitzpatrick, Eleonora Patacchini, and Nicolas Ziebarth, who offered me many useful comments that have greatly improved my dissertation. Moreover, I am grateful for John Forester, the then DGS of Cornell's CRP program, who made me choose Cornell as my PhD school. I have also received suggestions from my classmates and other PhD students at Cornell, including Danny Adiwibowo, Omar Ali, George Berry, Marisa Carlos, Pinyi Chen, Christine Coyer, Becca Jablonski, Woosung Kim, Daniel Kuhlmann, Sneha Kumar, Yanan Li, Lisha Liu, Yuanyuan Liu, Yuqi Lu, Hang Lü, Maricar Mabutas, Alberto Morales, Fernando Plascencia, Emily Taylor Poppe, Mauricio Sarrias, Meicheng Wang, Tao Wang, Youngmin Yi, Sherry Zhang, Ziye Zhang, as well as many others.

During my doctoral studies I have presented my research at numerous seminars and conferences, and I have been fortunate to receive comments from professors at other institutions, as well as journal editors and referees who have reviewed my articles over the past few years. I would especially like to thank Jere Behrman, Sandra Black, George Borjas, Francesca Cornaglia, Matz Dahlberg, Esther Duflo, Gilles Duranton, V. Jeffrey Evans, Ryan Finnigan, Alfonso Flores-Lagunes, Rachel Franklin, Edward Glaeser, Laurent Gobillon, Limor Golan, Clément Imbert, Yannis Ioannides, Lawrence Katz, John List, Janice Madden, Alan Manning, Paulo Masella, Ted Mouw, Ryan Muldoon, Çağlar Özden, Mallesh Pai, Robert Pollak, Hillel Rapoport, Tony Smith, Erdal Tekin, Brigitte Waldorf, Michael Weisberg, Xiaoyu Xia, Yuichi Yamamoto, Yves Zenou, and many other scholars who have made my research much better. As a non-economist working on a variety of applied microeconomic topics, I have been extremely lucky to receive many valuable suggestions from mathematics, statistics, and economics PhDs outside Cornell, including Xuan Bi, Shu Cai, Kilian Heilmann, Ara Jo, Cong Liu, Maryam Naghsh Nejad, Yumeng Ou, Matteo Sandi, Shihan Shen, Ruoying Wang, Yuyao Wang, Zhaoning Wang, Zhiling Wang, and Zoey Yi Zhao.

I am also thankful for financial support from Cornell University, Mario Einaudi Center for International Studies, London School of Economics, European University Institute, Royal Economic Society, as well as my parents, who supported my travel when conference grants were (usually) small.

I owe special thanks to my family and friends. First, this doctoral dissertation is meaningful and important to my parents. Studying for a PhD abroad has been an exciting individual challenge, but they have also shared it and considered it to be a part of their lives.

I would like to thank Jing Zhang, for always having high expectations of me as well as my dissertation. During my PhD studies I frequently traveled to her cities—Cambridge and Beijing—to conduct various research projects, and she always accommodated me with care.

I would also like to thank Jiabin Liu and Shihan Shen, who are my life-long friends, really nice guys, and (un)fortunately also PhDs who have always known my feelings. They kindly shared some of the best and worst moments during my PhD studies.

I am also thankful for my friend Yunjie Li, for providing me with many interesting and insightful discussions when I went through a difficult time in my second year.

Finally, I would like to dedicate this dissertation to my piglet, Zhuzhu the Arctic, who does not speak human language, but has always patiently listened to my stories, thoughts, and ideas, and never judged. You might not understand how stressful this three-year journey[1] is, so you might never know how helpful you are.

[1] In fact, similar to many other PhDs (for example, John von Neumann!), it has been relatively relaxing for me to be a researcher, but think about being *Homo sapiens* at the same time! This is much more complicated, exhausting, and difficult. Therefore, I would still thank you even if I had never started my PhD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation analyzes social networks in the context of international migration. Although social network research is not a traditional field of social science, there has been a burgeoning literature on social networks in recent decades. However, there are still a variety of unexplored topics in this newly developed field. This dissertation attempts to fill several crucial gaps in the field of social network research and contribute to the existing literature along three dimensions. First, this dissertation studies two important—yet less explored—research questions: social network formation and social network characteristics. Specifically, it studies why and how individuals form the social network based on their needs and interests, and how the social network can be defined and further observed based on the representative characteristics of the network. Second, this dissertation studies individuals' social networks in contexts of *intersectionality* of the population. Different from early social network studies that put main focus on the general population, this dissertation mainly studies specific immigrant populations that are associated with certain educational background, human capital characteristics, or professions. By doing so, it highlights the heterogeneity of social networks among various populations. Finally, this dissertation attempts to fix the econometric issues widely seen in traditional social network research. Specifically, it examines how individuals' needs and interests cause network formation, how individual characteristics are causally related to the representative characteristics of the network, and how the network leads to individual behaviors and decisions. In other words, the dissertation studies network formation, network characteristics, and network effects in a causal manner.

Before discussing any theoretical and methodological details of this dissertation, it is important to review the origin of social network research. Economists have long studied how the market responds to political events, social changes, policy shocks, etc., and furthermore, how individuals' social and economic behaviors are affected. Since Granovetter's pioneering research on "weak ties" (1973), however, social scientists have started to observed that individuals' networks of social relationships can also generate various types of "non-market" effects (Glaeser and Schienkman, 2001), and lead to social and economic consequences at both the micro and macro level.

Scholars in different disciplines study social networks from different perspectives. For example, sociologists mainly rely on social theories to predict social network formation, analyze the channels and mechanisms through which social networks affect individuals, and further investigate the effects of social networks (e.g., Granovetter, 1973; McPherson et al., 2001). Theoretical economists, on the other hand, mainly focus on the assumption that individuals form social networks by maximizing their utility function (e.g., Jackson and Wolinsky, 1996), and after social networks are formed, how social networks benefit individuals (e.g., Montegomery, 1991). Guided by these economic theories, empirical economists (e.g., Munshi, 2003; Calvó-Armengol et al., 2009; Damm, 2014) examine specific individual outcomes (such as wage, health behavior, and educational attainment) and econometrically evaluate the magnitude of the social network effect on these individual outcomes. In urban planning and geography, scholars also study social networks but primarily consider that networks are embedded in the geographic context. Therefore, early geographers mainly examine the "neighborhood effect" (e.g., Cox, 1969), which is later viewed as a special form of the social network effect. Such geographic research provides

valuable insights for traditional social network studies in sociology and economics, as people are usually geographically concentrated (in, e.g., residence, workplace—see Schelling, 1969) and interacting within spatial relationships is the fundamental channel through which social networks are formed (Marmaros and Sacerdote, 2006).

Although theories and methodologies of social network research have been developed for more than four decades, and many empirical results of social network research are reliable and have important policy implications, there are still a large number of topics unexplored in this broad field. Indeed, there are disproportionately many studies that attempt to examine and statistically measure social network effects, but there is still lack of empirical research on how social networks are formed and presented, and whether such findings can be consistent with existing theories that explain social network formation and the presence of network characteristics.

Another concern of traditional social network studies is that such studies focus too much on the general population, without taking its huge heterogeneity into careful consideration. The U.S. society has become increasingly more diverse in recent decades (e.g., Farley and Haaga, 2005; Lichter, 2013), and the U.S. is definitely not the only country that has experienced such social and demographic changes. Social networks might be formed differently in different populations, and further lead to different social and economic consequences.

This dissertation sheds light on social network research from the above perspectives. I conduct three empirical studies of social network formation, characteristics, and effects in three specific immigrant populations. This chapter presents a brief introduction before I explore the above topics in case studies.

1.1 Social Network: Its Formation, Characteristics, and Effects

Researchers have realized that individual behaviors and outcomes could be influenced by others even before the concept of *social network* was developed. In sociology of education, the Coleman Report (1966) points out that students' academic achievement is not only determined by teacher quality and family background, but also classmates' achievement. Cox (1969) observes the geographic concentration of political views and considers *neighborhood* as a crucial contextual variable. Since Granovetter's work (1973), researchers have started to formalize the concept of social network and study various network theories (e.g., Montegomery, 1991; Jackson and Wolinsky, 1996) and empirical considerations (e.g., Manski, 1993, 2000) to examine how much an individual can be affected by the social network—and the network is usually defined by researchers themselves.

Before any empirical analysis of the social network effect, however, it is important to figure out the following questions: what is a network? Who are network members? Who are expected to really influence others? Indeed, some social networks might have very complicated structures where there are "key players" and "outliers" (Jackson and Wolinsky, 1996; Calvó-Armengol et al., 2009; Lin and Weinberg, 2014) who affect others differently, not to mention that researchers might even include outsiders into the social network who do not have any influence (Foster, 2006; Stinebrickner and Stinebrickner, 2006).

Due to the above concerns, in recent years researchers have started to focus on a more fundamental question in social network research: how are social networks formed? People have had developed the answer of this question cen-

turies ago, as an old proverb in Medieval England says: "birds of a feather flock together". Such preferences of social network formation have largely been studied by sociologists who develop the literature of *homophily* (e.g., McPherson et al., 2001) to demonstrate that people associate and thus form their social networks with similar others. However, these descriptive findings still cannot answer the question: why are social networks formed based on that people are homophilous? Of course, a simple explanation of the presence of homophily in the social network is that individuals are happy when bonding with similar others. However, another possibility is that people form need-based social networks and peer effects within network are strong if network members share some similar characteristics (Marmaros and Sacerdote, 2006). For example, students' academic outcomes are determined not only by school quality and family background, but also classmates and schoolmates (e.g., Coleman, 1966). In this context, a specific example involves peer effects on learning that are generated among second-generation immigrant children of the same ethnic or cultural origin (e.g., Hoxby, 2000). Focusing on social network effects, this economic perspective adds to the traditional literature of social theories, and provides an alternative—and probably more important—explanation of network formation.

The above analysis of network formation leads to the next research question: if a social network is formed based on some criteria, attributes, or shared interests, then can we observe the representative characteristics of the network? This is also closely related to the way to *define* the social network. Traditionally, many social scientists have found various social networks formed solely based on similar demographic attributes, such as age, gender, skin, and most commonly, ethnicity (e.g., McPherson et al., 2001). It is easy to observe demographic-based networks but the mechanisms behind the emergence of network characteristics

can be either simple or complicated: such networks can be formed simply because individuals prefer to bond with similar others, but there might also be need-based mechanisms behind network formation. A typical example is the co-ethnic group of middle school students who hope to have better academic outcomes (e.g., Hoxby, 2000), but similar co-ethnic groups might exist even among individuals with very high levels of human capital accumulation, such as researchers (e.g., Borjas and Doran, 2012; Borjas et al., 2015; Freeman and Huang, 2015).

In any case, researchers have well explored the presence of the social network in which demographic attributes are the representative characteristics among network members. Less is known about the more general representative characteristics of social networks, which might be non-demographic. This is useful because many social networks are formed based on some behavioral characteristics *within* a specific demographic group. It is, however, much more challenging to study the behavioral characteristics of social networks, and one fundamental reason is that most recent social surveys (such as NLSY or AddHealth) do not have adequate information about behaviors of both respondents and their friends, even if the social networks are relatively well-defined. One exception is in the field of health economics: scholars have long observed the spread of health behaviors and physical attributes, such as diet preferences, exercise (Centola, 2011), and obesity (Christakis and Fowler, 2007; Trogdon et al., 2008). Compared with health characteristics, however, researchers pay much less attention on other representative characteristics of social networks.

The health economic literature further points out two possible channels through which a social network can be formed and, simultaneously, its rep-

resentative characteristics are presented. Specifically, the two channels are peer selection and peer influence. In the first channel, individuals form the social network by seeking others with the characteristics they like. In the second channel, individuals have already formed the social network based on some criteria, but network members influence each other and develop similar characteristics within the network. Clearly, the first channel—peer selection—is more related to the first question of this section, namely, network formation. Research on social networks formed based on demographic characteristics can only explore the channel of peer selection, as individual demographic characteristics cannot be influenced in most cases. The second channel—peer influence—plays its role in some other situations in which individual characteristics can be changed. Both channels lead to strong policy implications: studies of peer selection point out how policy makers can get involved in the process of network formation, and studies of peer influence often highlight the "social multiplier" (e.g., Glaeser, 2003) as the consequences of being in social networks. However, the above two channels might dominate in different situations, based on the research question that we are interested in. In general, peer influence is most powerful in social networks that are closely relevant to *outcomes* such as earnings or employment (Montegomery, 1991). On the other hand, peer selection dominates the channel through which social networks are defined by *behavioral* characteristics (e.g., Cohen, 1977; Kandel, 1978; McPherson et al., 2001; Christakis and Fowler, 2007; Centola, 2011; Shaefer and Simpkins, 2013), and in such situations the contribution of peer influence is usually overestimated[1] (e.g., Cohen-Cole and Fletcher,

[1] A possible explanation is that many types of labor market outcomes, which are influenced by other network members, are not related to any change in individual behavior and can be affected *immediately*. For example, individuals introduce job opportunities to others through the social network; by doing so, individuals' labor force participation, employment status, and even earnings can thus be changed once the social network starts to play its role (e.g., Munshi, 2003). These differences lead to the difference in the way individuals interact with the social network.

2008; Aral et al., 2009).

In the previous part of this section I have discussed how social networks are formed and how they can be defined by representative characteristics. Now I turn to the final question: what are the social and economic consequences of being in a social network, or more simply, what are social network effects? Although this is the last question presented in this section, it is probably most explored and has been widely studied in various disciplines no later than the 1960s (e.g., Coleman, 1966; Cox, 1969; Granovetter, 1973). A more challenging question is to understand the mechanisms behind the network effect.

On one hand, the investigation of social network effects is similar to the previous discussion of the channel of peer influence through which the social network is formed and defined. For example, social networks affect health outcomes because individuals are influenced by others and thus change their health behaviors. As discussed earlier, however, in many situations individuals in the network need not to experience any transition of behavioral patterns or characteristics in order to achieve the outcome, and the outcome can be directly affected by other individuals in the network. Such examples can be widely found in labor markets where social networks have "non-market" effects (Glaeser and Schienkman, 2001) even in situations without any market failure (e.g., Munshi, 2003). Although a social network in the labor market might affect individuals' ability and non-cognitive skills, and thus affect labor market outcomes gradually, the dominating possibility is that the social network has immediate effects on individuals' labor market outcomes—without changing any individual characteristics other than these outcomes—through various mechanisms, such as information transmission (e.g., Bikhchandani et al., 1992) and informal job re-

ferrals (e.g., Montgomery, 1991). Such *immediate effects* are with respect to either employers[2] or employees[3], but in both cases, we might not observe any significant change in the skill, ability, or behavioral characteristics of both the network members and the network itself.

1.2 Econometric Analysis: How Can We Empirically Examine These Questions (in the Quantitative Manner)?

1.2.1 Empirical Strategies

A social network can be empirically studied using various methods. In this section, I focus on the quantitative approach to statistically examine the social network, including its formation, representative characteristics, and effects on network members.

I first discuss the empirical specification of network formation. Let O_i^f be the "outcome" of network formation for an individual i. This outcome can be as simple as the size of the network, but can also reflect the network type, structure, or other complicated outcomes. Let F_i be i's specific individual characteristic based on which he attempts to form the network, then we have the basic specification as follows:

$$O_i^f = \alpha_0 + \alpha_1 F_i + \mathbf{X}_i \alpha_2 + \varepsilon_i \tag{1.1}$$

[2]For example, an individual gets a job through job referral because the employer knows his information through other persons in the same network as him (e.g., Granovetter, 1973; Montgomery, 1991; Munshi, 2003; Damm, 2009).

[3]For example, an individual's decision-making process is affected by the social network, and outcomes such as labor force participation and occupation selection are thus affected (e.g., Granovetter, 1985; Mizruchi and Stearns, 2001).

where $\mathbf{X}_i$ is the vector of control variables and ε_i is the error term. For simplicity I only consider the case that O_i^f is the network size. Therefore, this equation shows how the individual characteristic F_i is related to the number of members of the social network formed based on F_i, and the magnitude of network formation (given F_i) is reflected by α_1 estimated using OLS. Note that, however, these O_i^f members in this network need not to have the similar characteristic with i. For example, this OLS specification might be used to investigate how an immigrant make friends with the native if he tends to culturally assimilate and thus have some acculturational characteristics, reflected by F_i. This equation thus examines how the friendship with the native is formed, but clearly the native people in his network do not have such acculturational characteristics.

I now turn to the empirical investigation of the representative characteristics of the social network. Clearly, the prerequisite for this research question is to know whether the network is given. In general, there are two typical cases introduced as follows.

First, if the network is not given, i.e., whether individuals form the social network is unknown, then the task is to first figure out the existence of the network, and then determine whether there are any significant representative characteristics. Using the language of machine learning, this requires the approach of *unsupervised learning*, such as cluster analysis. The basic cluster analysis method, k-means, transfers this question into the following form: we want to partition n individuals into k sets, namely $\{S_1, S_2, , S_k\}$, given that n individuals are with characteristics $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ (the F_i in Equation 1.1 is an element of $\mathbf{X}_i$), where each individual has d types of characteristics[4]. The social networks can thus be determined based on one or more similar characteristics among individuals

[4]Mathematically, the vector $\mathbf{X}_i$ is thus d-dimensional.

(i.e.,potential network members) by the following objective function:

$$\underset{S_1,S_2,\cdots,S_k}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{X}\in S_i} \|\mathbf{x} - \mu_i\|^2 \tag{1.2}$$

where μ_i is the mean of all points in the point set S_i. In other words, this function finds the summation of least squares similar to what the OLS does in the regression analysis. After the networks are defined based on specific characteristics, we can determine whether these characteristics are truly representative based on our criteria. Note that there should be no standard answer to whether the representative characteristics are presented or not: although the above k-means equation is unrelated to the specific context of the empirical question, the "distance" is related to the context.

A typical example of investigating the representative characteristics of networks is Garip's research (2012) on the dynamics of Mexico-U.S. migration. She uses exactly the same approach and observe that there are several major groups of Mexico-U.S. migrants with distinct personal characteristics, such as age, gender, and educational attainment. Note that this method requires massive subsequent work on the context, in the sense that researchers need to explain why individuals with specific characteristics can be grouped into a network or a category.

If the network is given, i.e., individuals identify their network members, then this research question turns to estimate how representative these characteristics are. This requires *supervised learning*—in which social scientists are most familiar with the regression analysis. Now if there is one (potentially) representative characteristic, F_i, and the outcome variable O_i^c measures this characteristic

among i's network members, then we have the following specification:

$$O_i^c = \beta_0 + \beta_1 F_i + \mathbf{X}_i \beta_2 + \epsilon_i \tag{1.3}$$

where $\mathbf{X}_i$ is the vector of control variables and ϵ_i is the error term. For example, O_i^c can be the proportion of network members who share the same or similar characteristic with i, or the number of such network members. Therefore, β_1, estimated using OLS, represents how much individual characteristics are related to network members' characteristics. Although the above equation is very similar to Equation 1.1, the economic and sociological reasonings behind two equations are different. Here, Equation 1.3 should be used to test the theory which predicts that members of the network should share (or do not share) some representative characteristics, while Equation 1.1 is concerning how a social network is formed based on individual characteristics or interests.

I finally turn to the empirical investigation of social network effects. While in Equation 1.1 and 1.3 I focus on how individual characteristics lead to network outcomes, now I look at how the network affects individual outcomes. In this section I present a simple specification. Let $\tilde{O}_i^e$ be the individual outcome, such as earnings, we have:

$$\tilde{O}_i^e = \gamma_0 + \gamma_1 N_i + \mathbf{X}_i \gamma_2 + e_i \tag{1.4}$$

where N_i is the network variable, $\mathbf{X}_i$ is the vector of control variables, and e_i is the error term. In the traditional theoretical framework, especially in education and health, N_i represents characteristics or behaviors of other network members (e.g., Granovetter, 1985; Manski, 1993). However, N_i might also be other types of network variables, such as the network size (e.g., Munshi, 2003). Whether N_i represents network characteristics or network size depends on the

research question we are interested in, i.e., it is about the choice between studying network size versus network quality. In any case, γ_1 is the OLS estimate of the network effect on the individual outcome $\tilde{O}_i^e$.

1.2.2 Statistical Issues

I now discuss several statistical issues of the above approaches. I mainly focus on three regression equations introduced earlier, as social scientists generally assume that networks are given; moreover, mathematically, if the partition of individuals is the same as the presence of networks, then the k-means result should be equivalent to the OLS result, as they minimize the same objective function based on same criteria.

The main concern of the above regression equations is that both individual characteristics and the social network are usually endogenous. The major source of endogeneity widely discussed in social science is the "reflection problem" (Manski, 1993): individual outcome $\tilde{O}_i^e$ might also reversely affect outcomes of network members N_i. This issue of reversal causality still exists if we consider N_i as the network size, instead of outcomes of network members (e.g., Edin et al., 2003). This problem might similarly exist when investigating network formation and network characteristics. For research on network formation, for example, English proficiency might affect an immigrant's intermarriage—as a special form of social network; however, marrying a native speaker might reversely affect language skills (see the identification strategy of Bleakley and Chin, 2010). For research on network characteristics, for example, an individual might self select into the "obesity network" or influence others'

health variables; however, his own weight might also be affected simultaneously (see the context of Cohen-Cole and Fletcher, 2008).

A standard solution to the reflection problem is to find the reliable source of exogenous variation in the network variable. For example, network members can be determined by random control trial (e.g., Duflo and Saez, 2003), and thus network characteristics or outcomes are also randomized. Natural experiments created by, e.g., government and school policies, also exogenously determine network members (e.g., Edin et al., 2003; Stinebrickner and Stinebrickner, 2006; Damm, 2014), but there might be non-compliance cases and only the local average treatment effect (LATE) can be estimated. Researchers also rely on weather shocks that generate exogenous variation in networks (e.g., Munshi, 2003) and estimate the LATE. Note that if a natural experiment leads to the valid instrumental variable (IV), the LATE estimated using the IV approach is still unbiased, although the result should be interpreted with caution.

In most papers, solutions to the reflection problem focus on networks. However, if the independent variable is the individual characteristic—for example, if we investigate network formation—then it is similarly required to have exogenous variation in the individual characteristic (e.g., Bleakley and Chin, 2010).

I should also point out an additional issue of investigating the representative characteristics of the social network: the duo-channel problem. In Equation 1.3, even if we have a valid IV for the individual characteristic F_i, it is still difficult to separate out two types of effects, namely, peer influence and peer selection. Specifically, the IV approach solves the reflection problem that i is not influenced by network members; however, we still cannot determine whether the presence of representative characteristics of the social network is due to i's self-selection

or his influence on other network members. The unbiased estimate $\tilde{\beta}_i$ describes the overall magnitude of homophily, but the mechanisms can only be analyzed based on a theoretical framework, or other data structures (e.g., experimental[5], or longitudinal).

The reflection problem is not the only statistical issue. When using individual characteristics as independent variables to examine network formation and network characteristics, it is challenging to accurately measure these characteristics. Many related studies focus on the effects of either cognitive or non-cognitive skills on how networks are formed or presented. While some types of skills (such as years of schooling) can be accurately measured (e.g., Chiswick, 1980; Portes and Zhou, 1993; Mouw and Xie, 1999), measurement error might arise when studying other types of skills that cannot be accurately measured. Such unmeasurable skills mainly involve self-report cognitive or non-cognitive ability. For example, self-report language proficiency might be ambiguous since questions about language in social surveys are usually unclear or imprecise (Kominski, 1989). Measurement error leads to statistical bias, and the direction of bias might not be the same as that generated by the reflection problem. Hence, if both the reflection and measurement problem exist, it is even not possible to predict the sign of bias of the effect estimated using the OLS model (or similarly nonlinear model, such as logit).

The omitted variable issue, which is related to both individual and network variables, might further threaten the statistical analysis of networks (Manski, 1993). There are many types of unobservable factors that are related to both individual characteristics and the outcome variable of interest: for example, it is

[5]For example, there should be two randomized experiments: in one experiment individuals can both choose members of the social network and influence them, while in another experiment individuals cannot choose network members.

difficult to control (or fully control) for family background, attitudes, etc. Similarly, networks are also usually endogenous in the sense that many factors related to either network formation or network characteristics cannot be observed and included. In general, researchers can use similar methods that solve the reflection problem to fix both the measurement and omitted variable issue, but additional discussions must be made. For example, even if we find a valid instrument to tackle the reflection problem, it is still important to argue that this instrument is not relevant to the process of measurement and is unrelated to unobservable factors that might affect the outcome.

1.3 Social Networks and Immigration: Why Shall We Care about Social Networks among Immigrants?

I conclude this section by focusing on social networks and immigration, which will be the main topic of this dissertation. The U.S. is experiencing the so-called "third demographic transition" (e.g., Lichter, 2013), in which the immigrant population rapidly increases, and it is certainly not the only country that experiences this. Moreover, more than half of all countries in the world either send emigrants to other countries, or receive immigrants from other countries. Hence, immigration is a global issue related to both the developed and developing world.

It is important to study immigration from the social network perspective because, similar to any other population, immigrants form social networks with both immigrants themselves and the native, and such social networks can affect individual native and foreign-born people, as well as the local economy and so-

ciety. In general, a social network can generate benefits for its network members through various mechanisms. This is similar for ethnic social networks living in the host country: immigrants are more likely to offer job opportunities (e.g., Munshi, 2003), information (e.g., Damm, 2009), and generally any kind of informal help and support (Portes and Zhou, 1993) to other immigrants of the same ethnic origin. This has strong policy implications as immigrants are more likely to face disadvantages in the host society, such as discrimination (e.g., Oreopoulous, 2011; Rubinstein and Brenner, 2014), and receiving support from ethnic social networks improves immigrants' social and economic outcomes.

Furthermore, social networks among immigrants are not only related to immigrants themselves. The social concentration of immigrants might also affect the native-born, the local economy, as well as the society. The mass influx of immigrants might have impacts on the local labor market, and further affect at least some local populations. For example, the huge immigration of lowly educated foreign-born workers might hurt local low-skilled workers (Borjas, 2015; Peri and Yasenov, 2015). With the effects of ethnic social networks (i.e., the multiplier effects), such negative impacts on local workers might even be larger. On the other hand, researchers have long observed that the local economy is positively correlated with the presence of the foreign-born labor force and ethnic diversity. For example, ethnic diversity improves the productivity of both foreign-born and native workers (Peri, 2012) as well as the city-level economy (Ottaviano and Peri, 2005, 2006). Similarly, such positive effects of immigration become even larger if we take the effects of ethnic social networks into account.

Researchers study immigrants' social networks following the similar way of studying general social networks. In fact, social scientists started to investi-

gate ethnic social networks almost as early as early research on social networks: Schelling's theoretical framework (1969) of racial segregation mainly focuses on Black-White relationships in the U.S., but can be easily extrapolated to native-immigrant relationships in any country. The conclusion of his model implies the geographic concentration of individuals by ethnicity, which forms the spatial basis for ethnic social networks among immigrants. Indeed, empirical research on immigrants' locational choices shows that immigrants generally prefer ethnic enclave residence and choose to reside in areas where proportions of immigrants of the same ethnic origin are relatively high (Bartel, 1989; Altonji and Card, 1991; Portes and Zhou, 1993), and social networks among immigrants residing in ethnic enclaves are thus formed. Note that this process can be further influenced by individuals' assimilation characteristics, such as educational attainment and schooling (e.g., Bartel, 1989; Portes and Zhou, 1993): for example, immigrants who are more educated are less likely to choose ethnic enclave residence (Massey and Denton, 1985; Bartel, 1989; Bleakley and Chin, 2010). Also, it is worth mentioning that immigrants might also bond with other immigrants or the native and join social networks in the form of intermarriage (e.g., Gregory and Meng, 2005) or other types of social interactions (e.g., Bleakley and Chin, 2010).

Studies of network formation among immigrants point out race and ethnicity as the important representative characteristics of immigrants' social networks. While causal inference is traditionally challenging, in recent years scholars rely on natural experiments created by policies (e.g., random assignments of roommates conducted in colleges) and investigate the causal relationship between individual characteristics and network members' characteristics. These studies show that social networks among immigrants are indeed formed based

on selection on representative characteristics in the experimental setting as well: individuals are much more likely to form social networks with others who share similar demographic attributes (Marmaros and Sacerdote, 2006), and thus the representative characteristics of social networks are reflected by the demographic attributes (Mayer and Puller, 2008). Clearly, the presence of such representative characteristics can only be due to selection, as demographic characteristics usually cannot be influenced and changed.

The above studies highlight ethnic homophily when there are various single-ethnicity ethnic networks. A further question is: how about the social networks within these single-ethnicity ethnic networks? Indeed, immigrants definitely do not interact with every other individual of the same origin, and form social networks only with *some* compatriots. In such social networks, all network members share similar or same demographic attributes, but non-members might also have these attributes. In other words, here the demographic attributes are indeed representative, yet not the useful characteristics of the network. Although there have been a few empirical studies that focus on how social networks can be labeled by individual characteristics other than demographic attributes (e.g., Girard et al., 2015), researchers generally know much less about what kinds of non-demographic traits (such as personal preference or behavior) can reflect and define the social network.

Finally, researchers study the social network effects on a variety of immigrants' outcomes. Economists, geographers, and sociologists all put the main focus on the socioeconomic consequences of being in social networks for immigrants. Similar to empirical findings of general social networks, researchers also find the mixed effects of ethnic social networks. On one hand, immigrants

receive help and support from other members of ethnic networks, which lead to better educational (e.g., Hoxby, 2000; Friesen and Krauth, 2010) and labor market outcomes (e.g., Portes and Zhou, 1993; Edin, 2003; Munshi, 2003; Damm, 2009). However, it is worth mentioning that immigrants are not always better off when associating with others in ethnic social networks. The typical problem of immigrants' social networks is that discrimination against immigrants (e.g., Oreopoulous, 2011) might be worsened when immigrants stay in their ethnic social networks (e.g., Dustmann and Preston, 2003) and attempt to keep their own cultural identities, which is disliked by local residents (Battu and Zenou, 2010). Another problem is that staying in social networks retards immigrants' assimilation processes. In any case, ethnic social networks serve as the "social multipliers" (Glaeser et al, 2003) that strengthen the potential benefits for immigrants but are also likely to make their social problems worse. These factors thus generate mixed social network effects on immigrants' socioeconomic outcomes. The above findings highlight the fact that some immigrants might instead do not join any social network, or join interethnic networks rather than their own ethnic networks, and are likely to have better socioeconomic outcomes by doing so. The patterns of network choices depend on a large variety of individual characteristics, though, such as the ethnic origin (e.g., Portes and Zhou, 1993), educational attainment (e.g., Massey and Denton, 1985; Bartel, 1989), and language skills (Bleakley and Chin, 2010).

In the rest of this dissertation, I will conduct three case studies that concern the three topics that I discuss earlier in this section, namely network formation, network characteristics, and network effects. In Chapter 2, I study immigrants' carpooling behaviors when they commute to work. Labor economists and transportation scholars have found many determinants of immigrants' carpooling

behaviors (e.g., Teal, 1987; Ferguson, 1991; Huang et al., 2000; Charles and Kline, 2006; Cutler et al., 2008; Blumenberg and Smart, 2010). I further explore this topic by investigating immigrants' carpooling patterns from the network perspective: carpooling is not only an individual behavior, but also the channel through which immigrant carpoolers *form* a social network for commuting. I study the effect of English proficiency on the tendency of joining carpooling networks, and argue that such networks are need-based in the sense that immigrants who have less advanced English proficiency have to rely on other co-commuters more in order to tackle potential traffic difficulties, and thus English proficiency should be positively correlated with carpooling behaviors.

I then conduct a case study concerning network characteristics in Chapter 3. I focus on a special population in the single-ethnicity ethnic group, namely, Chinese graduate students in the U.S., and investigate the representative behavioral characteristics of their social networks. Specifically, I examine whether individual English-name usage is causally related to English-name usage in friendships. In the causal analysis, I attempt to exclude the possibility that individuals' English-name using behaviors are affected by friends, although individual English-name usage might lead to the network identity through both peer influence and peer selection, as argued earlier. This article studies network characteristics by focusing on individual behaviors: at the macro level, social networks are formed based on interactions among all network members; at the micro level, individuals actively shape the representative characteristics of the network by their name using behaviors.

In Chapter 4, I study how ethnic social networks affect migration outcomes of highly professional immigrants. Prior research mainly focuses on immigrant

workers who originally come from developing countries and acquire jobs in developed countries (e.g., Edin et al., 2003; Damm, 2009), and the labor markets are usually imperfect (e.g., Munshi, 2003). However, how about immigrant workers who migrate among developed countries and work in a highly globalized market? This case study focuses on French football[6] players who play in the English Premier Football League. Defining French teammates as members of the ethnic (French) social network, I study whether the network size affects immigrants' yearly migration outcomes—as football players usually discuss their future plans with football teams annually, there are three types of migration outcomes for French players: (a) staying in the current team; (b) staying in another team in England; (c) leave England. I will also discuss the heterogeneous network effects in specific subpopulations of French football players in England.

[6] As the paper is in the European context, I use *football* instead of *soccer* in the article. However, here I study the so-called *British football*, or *soccer*, instead of the football that is a popular sport in the North America.

CHAPTER 2

A STUDY OF THE FORMATION OF SOCIAL NETWORKS AMONG IMMIGRANTS: LANGUAGE PROFICIENCY AND CARPOOLING NETWORKS

2.1 Abstract of the Study

This study empirically examines the relationship between language proficiency and immigrants' carpooling behaviors. Using 2006 - 2010 American Community Survey (ACS) data, I focus on immigrants who were in childhood upon arrival in the U.S. and study whether English proficiency influences carpooling propensity and the number of co-riders, controlling for a large variety of demographic, socioeconomic, and geographic characteristics. To tackle the endogeneity problem of language proficiency and establish causality, I construct an instrumental variable (IV) for English proficiency based on the comparison of age at arrival between childhood immigrants who are originally from English-speaking and non-English-speaking countries. OLS and probit models as well as the IV regression model all show that individuals with higher levels of English proficiency are less likely to carpool and have fewer co-riders.

2.2 Introduction

Urban scholars have long observed that immigrants have special commuting patterns. For example, immigrants are more likely to carpool than natives (e.g., Cutler et al., 2008; Cline et al., 2009; Blumenberg and Smart, 2010, 2013).

There have been many theoretical carpooling models (e.g., Lee, 1984; Huang et al., 2000) and empirical studies of the determinants of carpooling, pointing out that gender (Crane, 2007; Buliung et al., 2009), race (Charles and Kline, 2006), education (Ferguson, 1997), occupation (Ferguson, 1991; Buliung et al., 2012; Jun, 2012), family (Adler and Adler, 1984; Hao, 2004; Blumenberg and Smart, 2010), housing and residential environments (Beckhusen et al., 2013), and region-specific characteristics (Teal, 1987) are all possible factors affecting carpooling behaviors. This article studies carpooling behaviors from the perspective of cultural assimilation (Gordon, 1964). Following prior research (e.g., Lazear, 1995; Alba and Nee, 2005), in this study cultural assimilation is measured by language proficiency.

Language proficiency can indirectly affect immigrants' commuting behaviors through various channels[1]. However, controlling for various demographic and socioeconomic factors, does English proficiency has a "direct" effect on carpooling behaviors? Using 2006 - 2010 (5-Year) American Community Survey (ACS) data (Ruggles et al., 2010), I empirically test the causal effect of English proficiency on immigrants' commuting patterns, including the carpooling behavior and the number of co-riders.

In this article, I focus on "childhood immigrants" who were under age 18 upon arrival in the U.S. I use both the traditional linear (OLS) and non-linear (probit) model, as well as the instrumental variable (IV) model to examine the relationship between language and commuting among these childhood immigrants. The IV tackles the concern that language proficiency is endogenous

[1] As for carpooling, for example, language proficiency is associated with a variety of potential determinants of carpooling behavior, such as earnings (e.g., McManus et al., 1983; Tainer, 1988), education (e.g., Kao and Tienda, 1995), and family (e.g., Kulczycki and Lobo, 2004; Bleakley and Chin, 2010; Duncan and Trejo, 2011). English proficiency is also heterogeneous among immigrants (e.g., Espenshade and Fu, 1997).

(Chiswick and Miller, 1995) due to, e.g., measurement error (Kominski, 1989). I follow the standard approach (e.g., Bleakley and Chin, 2004, 2010; Guven and Islam, 2015) and construct an IV for English proficiency using the interaction of the age at arrival and language characteristics of the country of origin. Both traditional and IV models show that immigrants with higher levels of English proficiency are more likely to drive alone and, on average, have fewer co-riders.

This article adds to the literature of urban and regional studies along two dimensions. First, it links immigrants' commuting patterns with the process of assimilation and, based on this demographic perspective, explores the effect of English proficiency on commuting behaviors. Second, the IV model used in this article identifies the effect of English proficiency in an arguably causal manner.

The rest of this article is structured as follows. Section 2.3 introduces the background and discusses the possible mechanisms behind the language effect on carpooling behaviors. Section 2.4 discusses the data set and statistical models. Section 2.5 reports the findings. Section 2.6 concludes the article.

2.3 Background

In this section, I will introduce the background of this article. In the first subsection I present a brief literature review on factors affecting carpooling behaviors. I then focus specifically on immigrants in the U.S. and introduce their carpooling patterns. I conclude this section by discussing the possible mechanisms of how English proficiency affects immigrants' commuting patterns.

2.3.1 The Determinants of Carpooling

Why do people carpool? Most studies focus on the effect of geographic factors and individual characteristics on carpooling propensity (Huang et al., 2000). Transportation scholars are interested in region-specific characteristics that are closely related to carpooling behaviors. For example, the availability of high occupancy vehicle lanes (Giuliano et al., 1990) positively affects carpooling propensity. Heavy tolls also encourage commuters to carpool (Nyerges and Aguirre, 2011). In addition, individuals choose whether to carpool based on the size of the city and the residential location (Teal, 1987). These geographic factors affect carpooling propensity through their effects on the commuting cost.

Individual characteristics are equally, if not more, crucial in affecting carpoolers. Although earlier studies (e.g., Horowitz and Sheth, 1978; Teal, 1987) suggest that socioeconomic status does not distinguish between carpoolers and solo drivers, it is still useful to control for these variables due to their correlation with individuals' opinions towards the value of time (Becker, 1965), which will affect commuters' behaviors and choices (Huang et al., 2000). In fact, recent research points out that at least some socioeconomic variables might still matter. Higher education attainment accounts for the decline in carpooling in the U.S. in past decades (Ferguson, 1997). The effect of income is somewhat unclear: while some articles find no clear relationship between income and carpooling (Ferguson, 1997; Brownstone et al., 2012), survey findings show that carpooling is associated with lower income strata and saving money is a main purpose of carpooling (Correia and Viegas, 2011). Similarly, occupation or job-specific characteristics (such as workplace transport policies) are related to carpooling behaviors (Buliung et al., 2011).

Individual- and family-level demographic characteristics also affect carpooling behaviors. Age, gender, and race are all found correlated with carpooling behaviors (Brownstone and Golob, 1992; Crane, 2007; Buliung et al., 2009; Blumenberg and Smart, 2010). In addition, an individual's commuting behaviors affect (and are simultaneously affected by) behaviors of other household members (Bard, 1997), and carpooling propensity is increasing in the household size because multiple-occupant vehicle trips are more productive in larger households (Smart, 2010).

The above studies focus on why people carpool. On the other hand, why do people *not* carpool? The economic model of carpooling shows that carpooling propensity declines with the value of time (e.g., Huang et al., 2000), although the magnitude of the time value is somewhat debatable (Viton, 1992). In theory, since the unit time cost is non-negative, one purpose for driving alone (instead of carpooling) is to minimize the commuting time. Indeed, all else being equal, carpooling is associated with longer travel times to work due to pick-up/drop-off delay (e.g., Levinson and Kumar, 1994; Yang and Huang, 1999; Frank et al., 2008) and there are obviously additional time costs if passengers' destinations are not on the same route. This pattern can be similarly observed in the ACS sample: on average, solo drivers spend 24.659 minutes on commuting, while the average travel time to work for carpoolers is 29.408 minutes.

2.3.2 Immigrants' Commuting Patterns

I now turn to introduce immigrants' commuting patterns. Researchers have generally observed that immigrants are more likely to commute by carpooling

(e.g., Blumenberg and Smart, 2014) or using public transit (e.g., Cutler et al., 2008). However, there is still huge heterogeneity in carpooling propensity *within* the immigrant population. For example, those who live in ethnic enclaves are more likely to carpool and take public transit (Liu and Painter, 2012). The main explanation of immigrants' special carpooling patterns is that immigrants heavily rely on social networks (e.g., Boyd, 1989; Massey et al., 1993; Zhou and Bankston III, 1994; Sanders et al., 2002; Portes, 2010), and carpooling is a typical social behavior in networks, although there might be multiple mechanisms through which social networks affect immigrants' carpooling patterns.

It is, however, important to clarify that the term *social network* here does not only indicate networks of friends, schoolmates, or colleagues (which are general themes in contemporary network research). In the context of immigrants' carpooling, it is probably more precise to specify the social network as the *friendship network* and the *household network* separately: indeed, carpooling mainly happens among household members (e.g., Liu and Painter, 2012); however, although carpooling with neighbors or friends are usually unplanned and not always stable among immigrants (e.g., Shannon, 2016), external carpooling does exist (e.g., Adler and Adler, 1984; Charles and Kline, 2006).

Along with immigrants' special occupational distributions (Green, 1999) and locational choices of jobs (Preston et al., 1998; Blásquez et al., 2010), higher carpooling propensity might be a driving factor that accounts for the longer average commuting time among immigrants, as shown in Table 2.1: compared with natives, immigrants are both more likely to carpool and have longer travel time to work. Childhood immigrants who were under 18 upon arrival are still more likely to carpool (and have longer travel times to work), and there are about

20% of all non-citizen workers that are carpoolers.

Table 2.1: Individual Commuting Choices

	Driving Alone	Carpooling	Mean Travel Time	Observations
Native	85.48%	9.21%	25.409	3,901,967
Immigrants:				
All	75.88%	14.97%	28.877	566,466
Arriving ≤ 18	78.54%	13.45%	28.473	221,372
Non-citizen	71.29%	19.04%	28.303	206,259

2.3.3 The Role of Language in Shaping Commuting Patterns

Most of the related studies investigate immigrants' special commuting patterns by pointing out the difference between natives and immigrants. However, less is known about the heterogeneity *within* the immigrant population and how such heterogeneity creates variation in carpooling behaviors. In this article, I focus on a perspective that has not yet been well explored by prior research: the role of language. The effects of language proficiency on immigrants' outcomes have been well studied in labor economics. Immigrants who speak better English also have better labor market outcomes (e.g.,. McManus et al., 1983; Tainer, 1988), receive more education (e.g., Kao and Tienda, 1995), are more likely to marry natives (e.g., Duncan and Trejo, 2011), and are more spatially assimilated (e.g., Bleakley and Chin, 2010). There is also huge heterogeneity in language proficiency (e.g., Espenshade and Fu, 1997) and even strategies of English learning (e.g., Portes and Zhou, 1993) among immigrants with different ethnic backgrounds.

The above findings partly explain the role of language skills in shaping commuting patterns. But controlling for these channels, English proficiency might still have a *direct* effect. Sanchez et al. (2004) point out the language difficulty when immigrants use public transportation systems. The difficulty similarly exists when an immigrant drives alone: even with the car navigation systems, language skills might still be helpful when dealing with traffic emergencies (e.g., talking with police officers, confusion on the road). This provides the possible explanation of why an immigrant needs a social network when commuting to work, and how the size of this *carpooling network* affects his carpooling patterns. In theory, the need of having co-riders should be negatively correlated with language skills, and thus immigrants with higher levels of English proficiency should be less likely to carpool, and further have fewer co-riders.

2.4 Data and Empirical Strategies

This section introduces data and methods. I first discuss data and variables, and then the descriptive statistics. I finally introduce regression models employed in this study, including traditional linear and non-linear models, and the instrumental variable (IV) model.

2.4.1 Data and Variables

In this article, I use 2006 - 2010 American Community Survey (ACS) public-use micro data (Ruggles et al., 2010) to study English proficiency and carpooling behaviors among immigrants. ACS is a nationwide representative micro-level

survey data set, which surveys questions about immigration and ethnicity (such as "year of immigration" and "birthplace"), and it is easy to identify immigrant status in ACS. It provides an adequately large immigrant sample; however, I drop two types of immigrants in the empirical analysis: (a) I do not include individuals who are not in the labor force, are unemployed, or work from home, as this study focuses on patterns of *commuting to work*; (b) I only select childhood immigrants, i.e., those who were under 18 years old upon arrival in the U.S., because age at arrival is nearly not a choice variable for these childhood immigrants. After initial data cleansing and processing the sample size is 221,372.

In the regression analysis, I include two dependent variables about commuting patterns: the carpooling indicator, and the number of co-riders. Independent variables include the indicator of English proficiency, as well as a large set of explanatory variables that are found to be potential determinants of commuting patterns in prior empirical research. I now discuss these variables.

Dependent Variables

I first introduce two dependent variables of commuting patterns. Individuals in ACS need to specify their commuting modes, including driving (including driving alone and carpooling), taking public transit, walking, and others (e.g., using motorcycle). I can thus construct a carpooling indicator as the dependent variable. ACS further provides an ordered variable that reflects the number of co-riders (in ACS it is called "vehicle occupancy") in the car. Clearly, for individuals who drive alone, the value of this variable equals 1, and for carpoolers the value is at least 2.

English Proficiency

The key independent variable in this article is an indicator of English proficiency. In ACS there is a question about individuals' English skills including five options: (i) does not speak English; (ii) speaks only English; (iii) speaks English very well; (iv) speaks English well; (v) speaks English, but not well. For simplicity and robustness, in the regression analysis I construct a dummy variable measuring English proficiency in a coarse manner: this proficiency indicator E equals 1 if the individual answers: (ii) speaks only English, (iii) speaks English very well, or (d) speaks English well; it equals 0 otherwise (i.e., if choosing (i) or (v) in the survey). However, I also measure English proficiency in alternative ways: (a) I define "proficient" if the individual answers (ii) and (iii) in the question, i.e., if the individual speaks only English, or speak English "very well"; (b) I study commuters who speak only English. The empirical analysis based on these measures will be presented in main results as well as robustness checks.

One major problem of the proficiency variable is that it is usually difficult to precisely quantify levels of language proficiency in nationwide surveys (Kominski, 1989). Similar to many statistical issues in survey data (Bound et al., 2001), this implies that using English proficiency as the independent variable is associated with measurement error, a major source of the endogeneity problem. Hence the traditional Ordinary Least Squares (OLS) or probit regression model are unlikely to identify the causal relationship between language proficiency and commuting patterns. In the methodology section I will introduce the instrumental variable (IV) model to solve this problem.

Demographic Variables

ACS provides a large number of demographic variables. Similar to most survey or experimental data, ACS surveys individuals' age, gender, race, country of origin, marital status, citizenship, and household characteristics (including household size, number of children and number of children under 5 at the time of survey). In ACS individuals also need to specify the country of origin in the question of the birthplace, hence I am able to construct county-of-origin fixed effects.

Socioeconomic Variables

ACS surveys a large variety of socioeconomic variables. Although there is no consensus regarding whether there are significant effects on commuting patterns for many socioeconomic variables (e.g., income), it is still helpful to include these socioeconomic variables in the regression analysis if possible.

The socioeconomic variables used in the regression analysis include income (at both personal and family level), housing status (house value, homeownership, number of rooms and bedrooms, and age of structure), food stamp recipiency, and educational attainment. In addition, ACS provides individuals' occupation (based on SOC classification) information, which allows me to construct occupation fixed effects.

Geographic Variables

Individuals living in different regions in the U.S. might have different commuting patterns due to region-specific characteristics (e.g., the availability of public transportation systems, or the population density). In this article, I control for individuals' geographic information using county fixed effects based on state and county codes of residences provided by ACS. I also construct a set of metropolitan status fixed effects (e.g., whether the individual lives in the central city, the metropolitan area, or the non-metropolitan area).

2.4.2 Descriptive Statistics

I now report the descriptive statistics and present an overview of the above variables. I start with dependent and main independent variables in Table 2.2. The first panel of Table 2.2 describes two dependent variables. Conditional on driving, 14.6% of all immigrants are carpoolers, and the average number of commuters in the car is 1.218, while if focusing only on carpoolers, the average number of co-riders is 2.488. In other words, a large proportion of carpoolers commute with only one other companion. The next panel focuses on the main independent variable in the empirical analysis: 91.2% of all childhood immigrants in the sample are proficient in English.

The third panel turns to childhood immigrants' demographic characteristics. The mean age is about 28. On average, these childhood immigrants arrived at 10. 44.5% of childhood immigrants are female. 73.6% of immigrants in the sample are U.S. citizens. For ethnicity, I am able control for detailed race/ethnicity information but for simplicity here I only show the proportion of White, African

Table 2.2: Descriptive Statistics

	Mean	Std. dev.
Dependent Variables		
Carpooling†	0.146	(0.353)
Number of riders†	1.218	(0.676)
Number of riders‡	2.488	(1.111)
Proficiency		
English proficiency	0.912	(0.283)
Speak very good English	0.737	(0.440)
Speak only English	0.295	(0.456)
Demographic		
Age	28.473	(22.787)
Age at arrival	9.646	(6.078)
Female	0.445	(0.497)
Citizenship	0.736	(0.441)
White	0.545	(0.498)
African American	0.071	(0.257)
Hispanic	0.413	(0.492)
Married	0.635	(0.481)
Household size	3.716	(0.190)
Number of children	1.078	(1.252)
Number of children under 5	0.220	(0.522)
Socioeconomic & Geographic		
Personal income (log)	10.471	(1.069)
Housing value (log)	12.200	(0.898)
Homeownership (free & clear)	0.157	(0.364)
Number of rooms	6.360	(1.673)
Number of bedrooms	3.314	(0.917)
Food stamp recipiency	0.047	(0.211)
≤ high school/GED	0.393	(0.488)
Some college education	0.286	(0.452)
Non-metropolitan	0.069	(0.253)
Central city	0.170	(0.376)

Observations: 221,372, unless otherwise noted.

†: among drivers. Observations: 203,640.

‡: among carpoolers. Observations: 29,774.

American, and Hispanic immigrants. More than half of all childhood immigrants are White and more than 40% are Hispanic; only 7.1% are African American immigrants. Note that it is possible that some immigrants have multiple ethnic backgrounds. Nearly two-thirds of these immigrants are married. The mean household size is 3.716, and on average, there are multiple children in households.

In the fourth panel I present the descriptive statistics of socioeconomic and geographic variables. The average personal income (not reported here) is 55,595 dollars. Following the standard way I use the log income as the covariate. Similarly, I use the log housing value in regression models. 15.7% of immigrants in the sample own their homes "free and clear", while others do not (e.g., with mortgage or loan). The average number of rooms is 6.360 and almost half of the rooms are bedrooms. Food stamp recipiency is a widely used indicator of low income strata, which are possibly correlated with carpooling (Correia and Viegas, 2011). 7.1% of these immigrants receive food stamps. As for education, approximately 40% of immigrants in the sample finish no more than secondary education, while 28.6% receive some college education and the rest in the sample earn at least bachelor's degrees. The last panel concerns geographic characteristics: 6.9% of childhood immigrants live in non-metropolitan areas, and 17% live in central cities. The metropolitan status fixed effects, along with county fixed effects, control for immigrants' geographic information.

2.4.3 Empirical Strategies

To analyze the impact of English proficiency on commuting patterns, I first introduce the basic specification based on the linear model. Denote O_i as the outcome variable. I construct the Ordinary Least Squares (OLS) regression as follows:

$$O_i = \beta_0 + \beta_1 E_i + \beta_2 \mathbf{D}_i + \beta_3 \mathbf{S}_i + \beta_4 \mathbf{G}_i + e_i \tag{2.1}$$

where for individual i, E_i is the dummy variable indicating English proficiency. $\mathbf{D}_i, \mathbf{S}_i, \mathbf{G}_i$ are the vector of demographic, socioeconomic, and geographic characteristics, respectively. e_i is the error term. Non-linear models (such as probit and ordered probit model) for regressions where O_i is dichotomous or ordinal can be similarly established.

The challenging statistical problem of Equation 2.1 is that E_i is endogenous. In most research on the effect on English proficiency, reversal causality (O_i simultaneously affects E_i) is the main threat, although this should not be a serious problem in the context of this article. Nevertheless, the measurement error problem of English proficiency (Kominski, 1989; Bound et al., 2001) still exists.

In this article, I employ the standard econometric approach (Angrist et al., 1996; Sobel, 2000) to use an instrumental variable (IV) for English proficiency to solve its endogeneity problem. The IV influences and "controls" the endogenous variable E_i to affect the outcome variable. Two requirements for a valid IV are: (a) the IV must be closely correlated with the endogenous variable; (b) the IV can only affect the outcome only through its effect on the endogenous variable (i.e., the "exclusion restriction").

My IV strategy follows the methodology of Bleakley and Chin (2004, 2010)

and Guven and Islam (2015): the age at arrival predicts English proficiency as childhood immigrants are most likely to acquire native-like proficiency when arriving during the "critical period" (Lenneberg, 1967). Psychological research considers nine years old as the boundary of the critical period (Johnson and Newport, 1989), which I will follow in this study. The exclusion restriction of the IV, however, implies that "age at arrival" itself cannot be a valid IV for English proficiency because it can influence the outcome through cultural (other than language) channels. Indeed, childhood immigrants arriving in the U.S. earlier are also exposed to American culture earlier, and there can be the effect of culture—which differs from the effect of language—on commuting.

To solve this problem, the constructed IV is based on the comparison between childhood immigrants from English-speaking countries and those from non-English-speaking countries. Both groups of immigrants experience the process of cultural assimilation, but childhood immigrants from English-speaking countries have already had English proficiency upon arrival. This leads to the following parametrization:

$$IV_i = \max(0, a_i - 9) \times NE_i \tag{2.2}$$

where a_i is i's age of arrival, and NE_i indicates that the individual is originally from a non-English-speaking country. Figure 2.1 visualizes this parametrization. Note that in this figure the definition of *English proficiency* is based on that the immigrant can speak "very good" English (by answering "very well", or "only", but not "well").

I first pool the immigrant sample and observe the general pattern that levels of English proficiency decline with age at arrival. Specifically, for childhood immigrants from non-English-speaking countries, age at arrival has the signif-

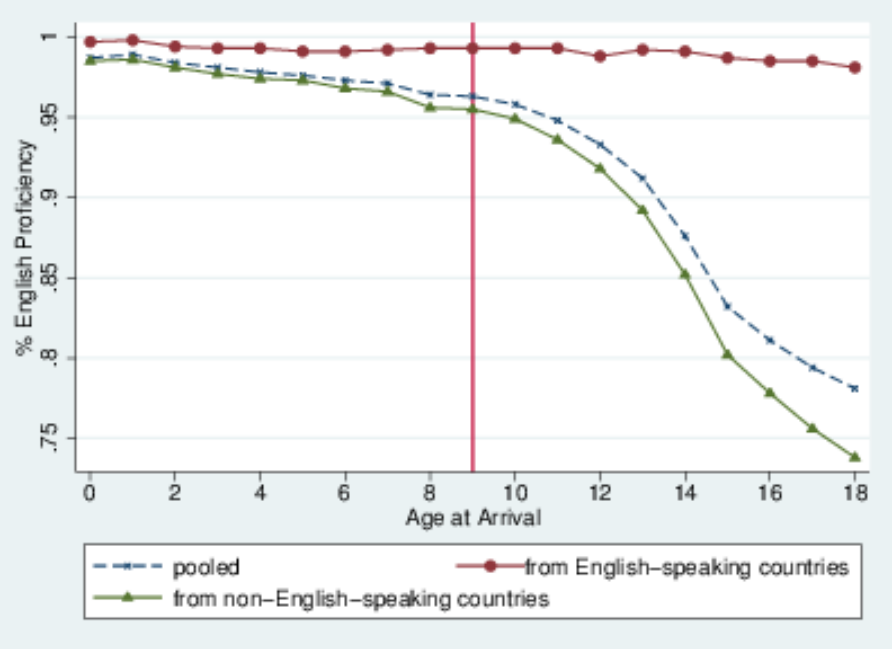


Figure 2.1: Age at Arrival and English Proficiency by Country of Origin

icant effect on English proficiency. In sharp contrast, age at arrival has almost no effect on English proficiency among childhood immigrants from English-speaking countries. This implies that the interaction between the age at arrival and the non-English-speaking-country origin controls for the cultural effect and influences immigrants' commuting patterns only through its effect on English proficiency. The above strategy leads to the first-stage regression:

$$E_i = \gamma_0 + \gamma_1 IV_i + \gamma_2 \mathbf{D}_i + \gamma_3 \mathbf{S}_i + \gamma_4 \mathbf{G}_i + \varepsilon_i \tag{2.3}$$

The fitted value of E_i obtained by Equation 2.3 is then used in Equation 2.1 to estimate the effect of English proficiency on commuting patterns.

2.5 Empirical Analysis: Carpooling Networks

This section reports the empirical findings of this article. I start with the regression analysis of English proficiency and carpooling. Table 2.3 examines the relationship between immigrants' English proficiency and the choice of carpooling using four regression models, including two probit models, one OLS model, and finally the IV model.

The first regression model, presented in Column 1 and 2, is the probit regression of the carpooling choice on English proficiency and other explanatory variables. I include all explanatory variables as well as race fixed effects, country-of-origin fixed effects, county fixed effects, and occupation fixed effects. Results show that English proficiency is indeed negatively correlated with the carpooling indicator. This is followed by the second probit model, in which I drop observations choosing commuting modes other than driving (e.g., taking public transit), and keep solo drivers and carpoolers only. The similar pattern is again observed: in general, childhood immigrants with higher levels of English proficiency are less likely to commute to work by carpooling.

I then employ the third model, i.e., OLS-based linear probability model to examine the size of the effect of English proficiency. The result can be interpreted in the linear manner: holding all other factors constant, an immigrant with English proficiency is 4% less likely to carpool than an immigrant with limited English proficiency. The effect, albeit statistically significant, is fairly small. In the remainder of Table 2.3 I run the fourth model, namely, the IV regression, to examine the effect of English proficiency on the choice of carpooling. The IV result shows that the OLS estimate is downward biased due to the endogeneity

Table 2.3: English Proficiency and Carpooling

	Probit		Probit		OLS		IV	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
English Proficiency	−0.108***	(0.012)	−0.129***	(0.013)	−0.040***	(0.003)	−0.118***	(0.018)
Age	−0.005***	(0.000)	−0.006***	(0.000)	−0.001***	(0.000)	−0.001***	(0.000)
Female	0.088***	(0.009)	0.091***	(0.009)	0.019***	(0.002)	−0.017***	(0.002)
Citizenship	−0.084***	(0.009)	−0.095***	(0.009)	−0.024***	(0.002)	−0.017***	(0.002)
Hispanic	0.083**	(0.026)	0.080**	(0.026)	0.014*	(0.005)	0.013*	(0.005)
Married	0.082***	(0.009)	0.071***	(0.009)	0.012***	(0.002)	0.010***	(0.002)
Household size	0.033***	(0.003)	0.036***	(0.003)	0.010***	(0.001)	0.009***	(0.001)
Number of children	−0.012**	(0.004)	−0.018***	(0.004)	−0.007***	(0.001)	−0.007***	(0.001)
Log personal income	−0.066***	(0.008)	−0.076***	(0.005)	−0.017***	(0.001)	−0.017***	(0.001)
Food stamp recipiency	0.051**	(0.016)	0.064***	(0.016)	0.016***	(0.004)	0.015***	(0.004)
Some college	−0.064***	(0.010)	−0.067***	(0.010)	−0.015***	(0.002)	−0.010***	(0.002)
College degree & above	−0.095***	(0.012)	−0.084***	(0.013)	−0.016***	(0.003)	−0.013***	(0.002)
Constant	0.486	(0.599)	0.037	(0.609)	0.396	(1.95*e*3)	1.204**	(0.351)
Full set of controls	Yes		Yes		Yes		Yes	
Race & origin FE	Yes		Yes		Yes		Yes	
County FE	Yes		Yes		Yes		Yes	
Occupation FE	Yes		Yes		Yes		Yes	
Log likelihood	−82,278.005		−79,524.652		—		—	
Pseudo R^2/R^2	0.058		0.060		0.054		—	
Observations	221,372		203,640		203,640		203,640	

e is "times ten raised to the power of". For example, "*MeN*" represents "$M \times 10^N$". *: $p < .05$; **: $p < .01$; ***: $p < .001$.

problem possibly caused by measurement error of language skills, and the IV regression estimates that all else being equal, the acquisition of English proficiency decreases the probability of carpooling by 11.8%.

In Table 2.3, I also present the effects of selected explanatory variables estimated by four models. Similar to the findings by prior empirical research, Table 2.3 shows that younger immigrants and female immigrants are more likely to

carpool (e.g., Brownstone and Golob, 1992; Crane, 2007), as well as those becoming U.S. citizens. Hispanic immigrants are more likely to carpool, which is also consistent with results in regional studies (e.g., Cline, 2009). Married immigrants and those in larger households are more likely to carpool, which can be explained by the strong "household effect" among immigrants (e.g., Liu and Painter, 2012), but carpooling propensity declines with the number of children. Some socioeconomic factors also influence immigrants' carpooling behaviors: personal income is negatively correlated with carpooling behaviors, and food stamp recipiency is positively correlated with carpooling behaviors. In addition, immigrants with at least some college education are less likely to carpool. This conclusion holds even after controlling for occupations. Finally, living in central cities increases carpooling propensity. These patterns are consistent with most prior findings: there are indeed demographic and geographic disparities in carpooling behaviors among immigrants, and some (although not all) socioeconomic variables are also related to carpooling behaviors.

Table 2.4 further examines the effect of English proficiency on the number of co-riders among immigrant drivers. Note that the number of "co-riders" for solo drivers is one. I now estimate the effect using two ordered probit models, then the OLS, and finally followed by the IV model in this table. In the first ordered probit model I focus on all immigrant drivers, and I find the negative association between English proficiency and the number of co-riders. In the second ordered probit model I only keep carpoolers (i.e., solo drivers are not included), and the qualitative pattern that the number of co-riders decline with language skills remains. The third model repeats the exercise of the first probit model, but now using the OLS regression. The size of the effect estimated using OLS is 0.04, but similar to the linear probability model in Table 2.3, here the

OLS estimate is again downward biased: the IV regression shows that English proficiency decreases the number of co-riders for an immigrant by 0.2, holding everything else constant. This number is actually not small, if compared with the average number of co-riders reported in Table 2.2 (i.e., 1.218).

Table 2.4: English Proficiency and the Number of Co-Riders

	Ordered Probit		Ordered Probit		OLS		IV	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
English Proficiency	−0.128***	(0.012)	−0.100***	(0.024)	−0.040***	(0.003)	−0.201***	(0.035)
Age	−0.006***	(0.000)	−0.005***	(0.001)	−0.001***	(0.000)	−0.002***	(0.000)
Female	0.081***	(0.009)	−0.046*	(0.020)	0.023***	(0.004)	0.021***	(0.004)
Citizenship	−0.085***	(0.009)	−0.034	(0.020)	−0.030***	(0.004)	−0.019***	(0.005)
Hispanic	0.076**	(0.026)	−0.018	(0.062)	0.021*	(0.011)	0.018	(0.010)
Married	0.054***	(0.009)	0.136***	(0.021)	0.003	(0.004)	0.001	(0.003)
Household size	0.036***	(0.003)	0.033***	(0.005)	0.018***	(0.001)	0.017***	(0.001)
Number of children	−0.009*	(0.004)	0.072***	(0.008)	−0.004*	(0.002)	−0.005**	(0.002)
Log personal income	−0.067***	(0.005)	0.027*	(0.010)	−0.021***	(0.001)	−0.021***	(0.002)
Food stamp recipiency	0.066***	(0.016)	0.036	(0.020)	0.037***	(0.007)	0.035***	(0.007)
Some college	−0.051***	(0.010)	0.074*	(0.030)	−0.009*	(0.004)	−0.001	(0.004)
College degree above	−0.071***	(0.013)	0.111**	(0.041)	−0.012*	(0.005)	−0.006	(0.005)
Central city	0.021	(0.012)	−0.025	(0.026)	0.002	(0.005)	0.001	(0.005)
Constant	—		—		1.323	(3.76*e*3)	2.260**	(0.675)
Full set of controls	Yes		Yes		Yes		Yes	
Race & origin FE	Yes		Yes		Yes		Yes	
County FE	Yes		Yes		Yes		Yes	
Occupation FE	Yes		Yes		Yes		Yes	
Log likelihood	−107,405.441		−25,124.02		—		—	
Pseudo R^2/R^2	0.034		0.052		0.044		—	
Observations	203,640		29,774		203,640		203,640	

e is "times ten raised to the power of". For example, "*MeN*" represents "$M \times 10^N$". *: $p < .05$; **: $p < .01$; ***: $p < .001$.

One concern of the above main results is that the definition of English pro-

Table 2.5: English Proficiency in Various Measures (All Regressions are IV)

	Proficiency: speak only English/speak very well				Proficiency: speak only English			
	Whether carpool		Number of co-riders		Whether carpool		Number of co-riders	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
English Proficiency	−0.057***	(0.008)	−0.144***	(0.015)	−0.086***	(0.009)	−0.200***	(0.020)
Age	−0.001***	(0.000)	−0.002***	(0.000)	−0.001***	(0.000)	−0.001***	(0.000)
Female	0.017***	(0.002)	0.021***	(0.004)	0.018***	(0.002)	0.022***	(0.004)
Citizenship	−0.017***	(0.002)	0.002	(0.004)	−0.018***	(0.002)	−0.003	(0.004)
Hispanic	0.014**	(0.005)	0.010*	(0.004)	−0.007*	(0.003)	−0.037***	(0.008)
Married	0.011***	(0.002)	0.012**	(0.004)	0.012***	(0.002)	0.014**	(0.004)
Household size	0.007***	(0.001)	0.012***	(0.001)	0.006***	(0.001)	0.009***	(0.001)
Number of children	−0.007***	(0.001)	0.003	(0.002)	−0.006***	(0.001)	0.005**	(0.002)
Log personal income	−0.016***	(0.001)	−0.011***	(0.002)	−0.016***	(0.001)	−0.010***	(0.002)
Food stamp recipiency	0.004	(0.003)	−0.001	(0.006)	0.003	(0.003)	−0.011***	(0.007)
Some college	−0.008***	(0.002)	0.014**	(0.005)	−0.014***	(0.002)	0.001	(0.004)
College degree above	−0.006**	(0.002)	−0.007	(0.005)	−0.011***	(0.002)	−0.019***	(0.005)
Central city	−0.026	(0.004)	−0.109	(0.009)	−0.033	(0.004)	−0.118***	(0.009)
Constant	0.355***	(0.033)	1.156***	(0.049)	0.287***	(0.022)	1.160***	(0.050)
Full set of controls	Yes		Yes		Yes		Yes	
Race & origin FE	Yes		Yes		Yes		Yes	
County FE	Yes		Yes		Yes		Yes	
Occupation FE	Yes		Yes		Yes		Yes	
Observations	203,640		203,640		203,640		203,640	

e is "times ten raised to the power of". For example, "*MeN*" represents "$M \times 10^N$". *: $p < .05$; **: $p < .01$; ***: $p < .001$.

ficiency might be too broad, and there is not much variation in language skills. In Table 2.5 I check the robustness by redefining English proficiency. In the first two tests I rerun two IV regressions of the carpooling behavior and the number of co-riders, but now considering only those who speak English very well or speak only English as the "English proficiency". Still, I observe the similar qualitative pattern: English proficiency is negatively correlated with the tendency of

carpooling, and is negatively correlated with the number of co-riders. The magnitudes of the effects, unsurprisingly, change with the new definition of English proficiency.

In the next two tests of Table 2.5 I only consider those who speak only English as immigrants who have English proficiency and rerun the IV regressions. Again, I find that English proficiency is negatively correlated with both the carpooling tendency and the number of co-riders. These robustness checks imply that the relationship between language skills and carpooling behaviors is not qualitatively affected by how English proficiency is defined.

2.6 Conclusion

This article uses 2006 - 2010 American Community Survey data to examine the relationship between English proficiency and commuting patterns among immigrants. I focus on the carpooling behavior and the number of co-riders among childhood immigrants who were under 18 upon arrival in the U.S.

Prior research has found a large variety of determinants of carpooling, including macro-level factors such as urban forms (e.g., Teal, 1987; Giuliano et al., 1990) and fuel cost (Huang et al., 2000), and micro-level individual characteristics such as age, gender, race (e.g., Brownstone and Golob, 1992; Crane, 2007; Blumenberg and Smart, 2010; Liu and Painter, 2012), family (e.g., Bard, 1997), occupation (e.g., Buliung et al., 2011), and education (e.g., Ferguson, 1997). The effect of income is somewhat debatable (e.g., Ferguson, 1997; Correia and Viegas, 2011; Brownstone et al., 2012). English proficiency might have the indirect effect on carpooling propensity because language skills are largely associated

with the above characteristics. This article examines whether English proficiency still has some "direct" effect on carpooling after controlling for these variables.

A major contribution of this article is methodological: I employ instrumental variable (IV) strategy, along with traditional statistical models, to examine the effect of English proficiency. The IV strategy tackles the statistical concern that language proficiency is endogenous (e.g., Chiswick and Miller, 1995) and establishes causal relationship between English proficiency and commuting patterns. I follow the approach by Bleakley and Chin (2004, 2010) and Guven and Islam (2015) and use a constructed variable—the interaction between an immigrant's age at arrival and the indicator of the non-English-speaking-country origin—to instrument for English proficiency. This is based on the psychological finding that immigrants are most likely to acquire native-like language proficiency if arriving during the "critical period" (Lenneberg, 1967), and comparing immigrants from non-English-speaking countries with other immigrants from English-speaking countries separates the language effect from the cultural effect.

Both traditional regression models (probit and OLS) and the IV model show that childhood immigrants with higher levels of English proficiency are less likely to carpool, and have fewer co-riders. This conclusion is robust across various sub-samples. In particular, the IV model implies that all else being equal, having English proficiency decreases the likelihood of carpooling by 11.8% and the number of co-riders by 0.2 for an immigrant.

This article adds to the literature of urban and regional studies along two dimensions. First, this article considers the effect of English proficiency from

the assimilation perspective, and examines whether English proficiency affects commuting patterns among immigrants, other than the indirect effect through its association with a variety of socioeconomic outcomes. The conclusion explores the current understanding of the determinants of carpooling behaviors among immigrants. Second, this article acknowledges the statistical challenge of the identification of the language effects. By introducing the instrumental variable model, I provide not only correlation, but arguably also the causal conclusion about the relationship between English proficiency and carpooling behaviors.

CHAPTER 3

A STUDY OF THE CHARACTERISTICS OF SOCIAL NETWORKS AMONG IMMIGRANTS: ACCULTURATIONAL HOMOPHILY

3.1 Abstract of the Study

Economists have long recognized the influence of friends on various outcomes among immigrants, and also observed that acculturation benefits immigrants from developing countries. This paper lies at the intersection of the above two topics: by focusing on a typical behavior of acculturation, namely English-name usage, I examine the extent of acculturational homophily among Chinese students. Specifically, I identify the causal relationship between self English-name usage and English-name usage of close friends using online social networking data on students who receive undergraduate education in China and graduate education in the U.S. The analysis relies on a natural experiment: English-name usage is affected by the difficulty of pronouncing the Chinese name by native speakers of English. Results show that individual acculturational behaviors lead to acculturational homophily in friendships: students with English-name usage have more close friends who also use English names, and the effect of self English-name usage is not through the number of close friends overall.

3.2 Introduction

Economists have long recognized the huge influence of friends[1]. This is also true for the immigrant population: for example, ethnic social networks affect labor market (e.g., Munshi, 2003; Damm, 2009) and educational outcomes (e.g., Hoxby, 2000) of immigrants and minorities originally from developing countries. Sociologists propose the theory of homophily, i.e., the tendency that people bond with similar others (McPherson et al., 2001), to explain how friendships can be defined.

Development economists also study the effect of culture on immigrants' lives. As the first stage of assimilation (Gordon, 1964), cultural assimilation—or more simply acculturation—usually leads to further economic (e.g., Arai and Thoursie, 2009) and social assimilation (e.g., Bleakley and Chin, 2010).

This paper lies at the intersection of the above two topics. I consider a typical behavior of acculturation—English-name usage—and investigates the extent of *acculturational homophily* based on it. Specifically, I use social networking data on Chinese students in the U.S. to identify the relationship between self English-name usage and English-name usage of close friends. This paper adds to the literature of development economics and demographic economics by examining the social consequences of efforts for acculturation among immigrants from developing countries. It also sheds light on network economics by studying the relationship between individual behavior and group identity.

Estimating the causal association between self English-name usage and

[1]Some related studies focus on network effects on earnings and labor market outcomes (e.g., Montgomery, 1991; Marmaros and Sacerdote, 2002), test scores and grades (e.g., Zimmerman, 2003), behaviors (e.g., Gaviria and Raphael, 2001, Kremer and Levy, 2008), and financial support (e.g., Blumenstock et al., 2016).

English-name usage of friends faces two challenges. First, many surveys do not provide information about friends' characteristics. Second, English-name usage can be endogenous. The advantage of using social networking data is that it provides information about individuals' self-nominated friends, as well as their English-name usage. The empirical analysis relies on a language-based natural experiment: for an individual, the difficulty of pronouncing the Chinese name in English exogenously affects his English-name usage.

The main results of this paper can be summarized as follows. Acculturational homophily based on name usage appears to exist among Chinese students: Conditional on nominating nonzero close friends, a student with English-name usage has nearly one more close friend who uses an English name. Overall, the presence of homophily does not rely on the total number of close friends nominated, and the empirical results are robust across subpopulations and for alternative measures.

The rest of the paper is structured as follows. Section 3.3 introduces the background. Section 3.4 proposes a theoretical framework that explains the mechanisms behind acculturational homophily. Section 3.5 discusses data and methods. Section 3.6 reports the results. Section 3.7 concludes.

3.3 Background

This section introduces the background of this paper. In the first part of this section, I review the literature on local-name usage among immigrants. I also briefly discuss the trade-off of using local names, and analyze the potential consequences of homophily based on local-name usage. In the second part, I

focus specifically on English-name usage among Chinese students in the U.S.: I analyze the determinants of English-name usage and discuss the natural experiment on English-name usage.

3.3.1 English-Name Usage and Immigrants' Social and Economic Outcomes

Along with many measures such as language attainment (e.g., Gordon, 1964; Mouw and Xie, 1999), local-name usage and naming convergence with the native are widely used to measure efforts for acculturation (e.g., Abramitzky et al., 2014). Fryer and Levitt (2004) pioneer the economic analysis of name: they find that the use of the typical Black name is correlated with socioeconomic status. In particular, scholars find that renouncing foreign names improves immigrants' earnings (Arai and Thoursie, 2009).

Immigrants who use local names are more likely to *actively* develop characteristics that improve social and economic outcomes. Local-name usage helps immigrants overcome "cultural barriers", which are common in developed countries (Belot and Ederveen, 2012). It also motivates language learning and could affect language skills (e.g., Edwards, 2006), which further improve labor market outcomes (e.g., Chiswick and Miller, 2001; Berman et al., 2003). In addition, using local names reduces the possibility of being discriminated. Here immigrants *passively* receive benefits from local-name usage. Ethnic-sounding names are usually related to labor market discrimination (Rubinstein and Brenner, 2014). Bertrand and Mullainathan (2004) find that workers with "White names" receive more callbacks for job interviews. This is even true for skilled

immigrants (Oreopoulos, 2011), and unsurprisingly, name-based discrimination against immigrants happens not only in the labor market (e.g., Drydakis, 2012; Zussman, 2013).

On the other hand, if local-name usage is really so good, why do some migrants refuse to use local names, even only *first* names? This can be explained by the association between name and identity (Nicoll et al., 1986; Larkey et al., 1993; Edwards and Caballero, 2008). There is a trade-off between retaining own cultural identities (e.g., name) and improving socioeconomic outcomes through acculturation (Battu and Zenou, 2010), and some immigrants reject the dominant culture because by doing so, they are more likely to benefit from their ethnic networks (Battu et al., 2007). This also reduces the cost of acculturation as some immigrants find it unlikely to be accepted by the mainstream society even after being culturally assimilated (Portes and Zhou, 1993).

The above analysis discusses potential consequences of English-name usage. A further question is: if individual English-name usage is related to English-name usage in friendships, how can individual benefit from homophily based on it? In theory, there are two main channels through which homophily matters. First, peer effects on assimilation-related outcomes (such as language skills for everyday life) can be generated in such friendships, especially in friendships within ethnic groups (e.g., Hoxby, 2000). Second, information about assimilation-related activities might be disseminated faster in such friendships. Economists do observe information spillovers in social networks (e.g., Topa, 2001; Duflo and Saez, 2003; Cai et al., 2015; Bennett et al., 2016), and migrant students hoping to acculturate benefit from friendships in a similar way[2].

[2]For example, many U.S. graduate schools organize workshops on employment in the U.S. for international students who hope to work in the U.S. after program completion. Many students might neglect the announcements, but information can be transmitted faster among

In theory, homophily might occur through two mechanisms. Economists have long recognized peer effects in education (e.g., Sacerdote, 2001; Foster, 2006; Calvó-Armengol et al., 2009) and in the context of this paper, diffusion of English-name usage can occur in close social relationships such as friendships. Another possible mechanism is peer selection, i.e., students self select into the friendship with (or without) the characteristic of English-name usage. Both mechanisms are consistent with the benefit-cost analysis of English-name usage and acculturational homophily discussed earlier, and the presence of homophily can be due to a mixture of both, although studies in economics and sociology of education generally point out that peer selection may play the more important role (e.g., Cohen, 1977; Kandel, 1978; Aral et al., 2009).

The above analysis explains the presence of homophily, and also suggests that economists should care about homophily based on acculturational behaviors. Individuals who want to culturally assimilate might personally use English names and also find themselves in friendships in which people use English names. Hence, they benefit from both self English-name usage and homophily based on English-name usage. On the other hand, individuals who do not want to culturally assimilate might neither use English names nor become close friends with those who use English names, as they do not highly value the potential benefits of both individual English-name usage and homophily based on it.

students who want to stay and assimilate.

3.3.2 English-Name Usage among Chinese Students

I now focus on English-name usage among graduate students from China. I will first introduce possible determinants of English-name usage explored in prior related work, and then discuss the natural experiment on English-name usage.

As a result of the English education system in China, most students *have* English names in classroom. China has began to accept Western culture in recent decades, and adopting the English name is not rare even among Chinese people staying in China. However, there does exist variation in English-name *usage*, as students are not required to use English names outside the classroom. English education might influence English-name usage and the general willingness of acculturation (Gao et al., 2005).

Because of the role of English education, educational attainment is likely to be associated with English-name usage. For example, better colleges might hire more international faculty members, offer better language courses, and have higher requirements for language learning. To take school quality into account, I split Chinese colleges and U.S. graduate schools into three tiers based on rankings and reputation-based school alliances. Appendix B of this dissertation presents the details of school-tier classification.

Age at arrival is another crucial factor affecting acculturational efforts. Migrants who arrived earlier are exposed to the dominant culture earlier (Bleakley and Chin, 2004, 2010) and have higher language skills (Espinosa and Massey, 1997). Hence, international students who arrived earlier might be more likely to use English names.

In addition, both pre- and post-arrival geographic characteristics can affect

English-name usage. It is unsurprising that the degree of assimilation is related to the local racial makeup (Bleakley and Chin, 2010), but pre-arrival variables might also matter (Polavieja, 2015). For example, immigrants originally from foreign-culture-friendly regions might be more willing to culturally assimilate. Local socioeconomic characteristics are also likely to influence individual efforts for acculturation.

In this paper, I restrict the sample to students who arrived in the U.S. *straight from* undergraduate programs in China, and most students were *born* in the late 1980s and early 1990s. The concentration of birth years minimizes the impact of unobservable time trends. Similar to other research based on social networking data, I am able to control for a variety of variables provided by the website. In the data section I will discuss the sample, variables, and students' backgrounds in detail.

I finally introduce a linguistic factor associated with English-name usage: the difficulty of pronouncing the Chinese name by native speakers of English. The presence of the *pronunciation difficulty* leads to name mispronunciations, which can further cause inconvenience, embarrassment, and discomfort for both Chinese and English speakers[3]. For students with difficult-to-pronounce Chinese names, one solution is to use an English name along with the original Chinese name, which suggests the relationship between the pronunciation difficulty and English-name usage.

In theory, the pronunciation difficulty of the name should be randomly "assigned" in the sample and arguably unrelated to individual characteristics. This

[3]One interesting example involves the character *mèng*, which has the meaning of *dream* in Chinese. This character is widely used in female given names, but is usually pronounced as *men* by English speakers.

is because that (a) the "linguistic distance" (Crowley and Bowern, 2010) between Chinese and English is long, and (b) naming decisions were made more than two decades ago when China was still a relatively insular country, and most parents born in the 1960s had limited knowledge of English, hence Chinese parents were unlikely to consider the difficulty of pronouncing their children's Chinese names in English. In Appendix C of this dissertation I present the simple yet robust criteria to identify difficult-to-pronounce Chinese names based on differences in linguistic properties between Chinese and English.

The randomization of the pronunciation difficulty is not testable as we cannot enumerate all characteristics. However, in Section 3.5 I examine the difference in observable characteristics between those who have difficult-to-pronounce names and those who do not have difficult-to-pronounce names, and generally find no significant difference in either individual characteristics (e.g., gender) or geographic characteristics (e.g., local economic and demographic variables) between two groups of students.

Another potential concern is that the pronunciation difficulty might be related to migration plans. It can be a caveat if the pronunciation difficulty is more commonly seen among migrant students. In Appendix D of this dissertation, I examine 18 external student samples retrieved from China Center for Economic Research at Peking University, and Nanjing Foreign Language School. These samples present students' names and post-graduation locational outcomes (i.e., staying in China or continuing education abroad). I find that (a) there is no significant difference in the percentage of students with difficult-to-pronounce names between students who stay in China and move abroad; (b) the percentage of students with difficult-to-pronounce names in these samples is very close to

that in the sample of this paper. This suggests the unlikely selection of student migration based on name pronunciation. Details are presented in Appendix D as well as Section 3.5.

I should point out that although the proportion of difficult-to-pronounce names tends to be universal, the effect of the pronunciation difficulty on English-name usage might be weak among non-migrant students. Due to the long process of preparation for standardized tests, students usually need to decide whether to pursue graduate education abroad in the early stages of the undergraduate program[4]. Therefore, many non-migrant students make early decisions of stay, have less incentive to make efforts for future acculturation in the U.S., and finally do not use English names as they are very unlikely to be exposed to native speakers of English, regardless of the pronunciation of names in English.

3.4 Theoretical Considerations

In this section, I introduce a simple theoretical framework that explains why the presence of acculturational homophily are expected. I start with a baseline model and then propose a two-stage model in which every individual migrant has an unobservable type of the willingness of acculturation, and will choose whether to start the (costly) acculturation behavior based on his own type.

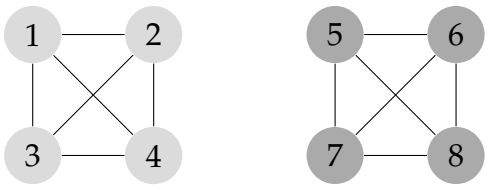
[4]This is particular true for Chinese students: before 2012, GRE tests (GRE/GMAT are required by arguably all U.S. schools) were only held two times per year in China, thus it is unlikely for a student who, say, decides to continue graduate education abroad after the junior year and starts school applications in the senior year, and manage to arrive in the U.S. right after college graduation.

3.4.1 The Baseline One-Stage Model

There are $2n$ individual immigrants who have either high (H) or low (L) type of the willingness of cultural assimilation; for each type there are n individuals. In this one-stage model, I assume the type can be observed and is consistent with (and thus precisely reflected by) some acculturational behavior. For example, an H-type individual uses the English name to reflect the willingness of cultural assimilation; an L-type individual, in contrast, does not use the English name.

For individual i, let $x_i = 1$ if i chooses to assimilate, i.e., i has the acculturational behavior, and $x_i = -1$ if i has not. For example, $x_i = 1$ if i uses an English name, and $x_i = -1$ otherwise. Again, in this model x_i is equivalent to i's type.

Social network effects of cultural assimilation can be generated within friendships among immigrants. In the first stage, these individuals form friendships, and friendship formation is costly. An example of friendship formation can be shown in the following undirected graph. In this graph, individual 1-4 are H-type (dark) and are friends, and individual 5-8 are L-type (light) and are friends.



Following the theoretical framework of friendship formation proposed by Jackson and Wolinsky (1996), I define the utility function $u_i(\mathbf{x})$ as follows:

$$u_i(\mathbf{x}) = x_i + \sum_{j \neq i} \delta^{t_{ij}} x_i x_j - f_{G_i} \tag{3.1}$$

before explaining these terms, I first introduce the denotations:

(1) $x_i = \pm 1$ indicates i's acculturational behavior (and thus i's type). G_i is the set of i's friends.

(2) t_{ij} is the graph distance between i and j: $t_{ij} = 0$ if i and j are friends (i.e., $j \in G_i$); $t_{ij} = 1$ if i and j are not friends (i.e., $j \notin G_i$), but i is the friend of j's one or more friends (i.e., $G_i \cap G_j \neq \emptyset$); etc. $t_{ij} = \infty$ can be similarly defined.

(3) δ is a distance discount factor and $0 < \delta < 1$.

(4) friendship formation is costly; for i, f_{G_i} is the total cost of friendship formation. Following the setting of Jackson and Wolinsky (1996) I assume the cost is identical regardless of the friend. Denote $0 < f < 1$ as the "unit cost" of friendship formation, then $f_{G_i} = f \cdot |G_i|$, where $|G_i|$ is the cardinality of G_i, i.e., the number of friends that i makes.

Now I explain the three terms in the above utility function. The first term, x_i, not only represents i's type and acculturational behavior; here, it also represents the utility i directly gets from the acculturational (or non-acculturational) behavior, holding all else constant. This term can reflect the reaction of the host society to i's acculturational behavior. Take the example of English-name usage: native residents appreciate that i uses the English name (i.e., $x_i = 1$), and i benefits from it; in contrast, not using the English name is not appreciated (i.e., $x_i = -1$), and i's utility decreases.

The second term, $\sum_{j \neq i} \delta^{t_{ij}} x_i x_j$, represents the utility i gets from his ethnic friendship. To start, let's consider the simple case that the multiplier $\delta^{t_{ij}}$ is not

introduced, and if i and j are friends, i gets utility $x_i x_j$ by making friend with j: only if both of them make the same choice of acculturation—in this context, both of them use (or do not use) the English name—can i gets positive utility from it, when $x_i x_j > 0$.

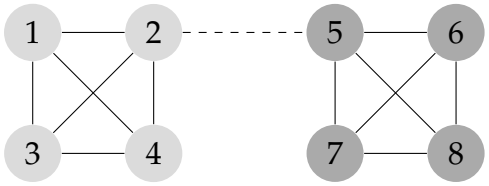
Now I assume that an individual can be affected by not only friends, but also friends' friends, friends' friends' friends, etc. I thus introduce the multiplier $\delta^{t_{ij}}$ ($0 < \delta < 1$) to describe the discount of the influence. Let t_{ij} be 0 if i and j are friends because in this case $\delta^{t_{ij}} = 1$. Then, let t_{ij} be 1 if i and j are not friends, but j is the friend of some i's friend, and in this case $\delta^{t_{ij}} = \delta < 1$, which reflects the discount of the influence. I further define $t_{ij} = \{0, 1, 2, \cdots, \infty\}$ as the graph distance between i and j, and $\delta^{t_{ij}}$ is decreasing as t_{ij} increases. $t_{ij} = \infty$ indicates that there is no path from i to j in the graph (as shown in the graph earlier, e.g., $t_{2,5} = \infty$ for individual 2 and 5), and $\delta^{t_{ij}} = \delta^{\infty} = 0$.

The final term, f_{G_i}, represents the cost of making friends within the ethnic group. As explained above, if making one friends costs f ($f > 0$), then the total cost of friendship formation $f_{G_i} = f \cdot |G_i|$, where $|G_i|$ is the number of i's friends. One might imagine that the cost of making friends with the different choices of acculturation (e.g., i makes friend with j, and only one of them is the English-name user) is higher, but later I will show that this does not change the result of the model.

If an individual maximizes his utility based on friendship formation, then this model will show that homophily based on acculturational behaviors occurs given certain conditions of f and δ. Formally, I have the following prepositions:

(1) If $x_i = 1$, then for all $j \neq i, x_j = 1 \Rightarrow x_j \in G_i$ if $f < 1 - \delta$; similarly, if $x_i = -1$, then for all $j \neq i, x_j = -1 \Rightarrow x_j \in G_i$ if $f < 1 - \delta$.

(2) If $x_i = 1$, then $j \in G_i \Rightarrow x_j = 1$; similarly, if $x_i = -1$, then $j \in G_i \Rightarrow x_j = -1$.



Simply put, (1) means that if $f < 1 - \delta$, then i and j must be friends if i and j have the same choice of acculturation (in the example of name usage, both i and j use the English name, or do not use the English name); (2) means that i and j must have the same choice of acculturation if i and j are friends, given any $f > 0$ and $0 < \delta < 1$.

Proof. (1) We only need to show that a sub-graph that contains all same-type individuals must be the complete graph. Consider the sub-graph G_{+1} in which $x_i = 1$ for all $i \in G_{+1}$; all of these n individuals are friends (i.e., linked). Now assume that $j, k \in G_{+1}$ are not worse off if they unfriend with each other. Following the utility function in Equation 3.1 this implies

$$1 + (n-1)1 \cdot 1 - (n-1)f \leq 1 + (n-2)1 \cdot 1 + \delta \cdot 1 \cdot 1 - (n-2)f \tag{3.2}$$

or more simply $1 \leq \delta + f$. If this is not true, i.e., $1 > \delta + f$, or $f < 1 - \delta$, then j, k should still be friends. This implies that G_{+1} is a complete graph if $f < 1 - \delta$ since j, k are two arbitrary individuals. Similarly, for the case of G_{-1} in which $x_i = -1$ for all $i \in G_{-1}$, G_{-1} is also a complete graph if $f < 1 - \delta$.

(2) This is clear: assume not (in other words, $x_i = 1$ but $x_j = -1$ and $j \in G_i$), then $x_i x_j < 0$ since $x_i = 1 \neq x_j$. In addition, for all $k \in G_j$, we have $\delta^{t_{ik}} x_i x_k < 0$ since $\delta > 0$ and $x_j = x_k$ (thus $x_i x_k < 0$). This indicates that i's utility will be lower if i makes friends with j.

Note that if the unit cost of making friends with the different acculturational behavior is higher, these results still hold: the result (1) is relevant if i and j have different acculturational behavior; in the result (2), f is an arbitrary positive number.

Under this simple setting, this baseline one-stage model predicts the presence of homophily in the case that the type (of the willingness of cultural assimilation) is consistent with and can be well observed through the acculturational behavior.

3.4.2 The Two-Stage Model

Now I consider a two-stage model. In the first stage, all individuals observe their own type, and need to make their own choice of acculturation (i.e., choose whether to culturally assimilate), but individuals do not interact with others and do not observe others' types. The second stage of the game is the same as the game introduced earlier: individuals decide friendship formation. This two-stage structure is similar to the theoretical framework of social networks and cultural assimilation proposed by Verdier and Zenou (2015), but without the consideration of education as a measure of assimilation.

The setting of the two-stage model is different from the previous one-stage model. In the previous model, an individual needs not to make his choice of acculturation, as it is perfectly reflected by his type. In this model, however, I relax the previous assumptions: in the first stage, the type of the willingness of cultural assimilation cannot be observed by others, and thus the type needs *not* to be consistent with the acculturational behavior. For example, in the previous one-stage model an H-type individual must assimilate, but in this two-stage model, an H-type person can choose not to assimilate, and an L-type individual can choose to assimilate.

I also assume that choosing to culturally assimilate is costly: for i, the cost is C_H if i is H-type, and the cost is C_L if i is L-type; $C_L > C_H > 0$, and in the second stage $x_i = 1$. If i chooses not to assimilate, there is no assimilation cost and in the second stage $x_i = -1$. This assimilation cost (C_H or C_L) represents all individual losses from choosing the acculturational behavior[5] and L-type individuals have the higher assimilation cost. However, this assimilation cost exists in the first stage and is thus not related to others' acculturational behaviors, which are observed in the second stage.

The above assumption implies that the utility function $u_i(\mathbf{x})$ in Equation 3.1 now needs to account for the assimilation cost C_{T_i}, where $T_i \in \{H, L\}$ is i's type.

Based on the above setting, an H-type individual has two options in the first stage: he either chooses to assimilate (do not imitate his type[6]), or chooses not to assimilate (imitate); similarly, an L-type individual has two options: he either chooses to assimilate (imitate), or choose not to assimilate (do not imitate).

[5]For example, it can be the administration fee, the utility loss of giving up own cultural identities, or any other cost incurred.

[6]The word "imitate" is also used in the original signaling game model (Spence, 1973).

Group Assimilation Behaviors of Two Types of Individuals

It would be complicated if there are multiple individuals of each type (i.e., $n > 1$) and players of the same type make different choices of acculturation. Now I study how individuals of the same type should make the same choice of acculturation.

Assume that there are k_H ($0 \leq k_H \leq n$) H-type individuals who choose to assimilate, and there are there are k_L ($0 \leq k_L \leq n$) L-type individuals who choose to assimilate. Assuming $f < 1 - \delta$, the utility of these k_L L-type individuals is (denoted as U_{L_1}):

$$u_{L_1} = 1 + (k_H + k_L - 1) \cdot (1 \cdot 1) - (k_H + k_L - 1)f - C_L = 1 + (k_H + k_L - 1)(1 - f) - C_L \quad (3.3)$$

Note that U_{L_1} is monotonically increasing in k_L, if holding k_H constant.

Denote $U_{L_{-1}}$ as the utility of L-type individuals who do not assimilate (i.e., do not imitate), then $U_{L_{-1}} = -1 + (2n - k_H - k_L - 1)(1 - f)$. $U_{L_{-1}}$ is a monotonically decreasing function of k_L. L-type individuals are indifferent in choosing whether to assimilate if only if $U_{L_1} = U_{L_{-1}}$, i.e., $k_H + k_L = n - \frac{2 - C_L}{2(1-f)}$. However, this equilibrium cannot unstable, as L-type individuals can either convert to assimilate or to not assimilate and then get higher utility.

As a result, the only stable cases are that all L-type individuals choose to assimilate, or all L-type individuals choose not to assimilate. Similarly, all H-type individuals should also make the same choice of acculturation. Hence, the $n \times n$ game can be converted into a 2×2 game.

Equilibrium

Now I study the equilibrium of the game of cultural assimilation. As discussed earlier, there are four possible situations in total, namely:

(1) H-type individuals assimilate, and L-type individuals assimilate. H-type individuals get utility $U_{H_A}(L_A)$; L-type individuals get utility $U_{L_A}(H_A)$.

(2) H-type individuals assimilate, and L-type individuals do not assimilate. H-type individuals get utility $U_{H_A}(L_N)$; L-type individuals get utility $U_{L_N}(H_A)$.

(3) H-type individuals do not assimilate, and L-type individuals assimilate. H-type individuals get utility $U_{H_N}(L_A)$; L-type individuals get utility $U_{L_A}(H_N)$.

(4) H-type individuals do not assimilate, and L-type individuals do not assimilate. H-type individuals get utility $U_{H_N}(L_N)$; L-type individuals get utility $U_{L_N}(H_N)$.

Or more simply, the following table describes the game of cultural assimilation, where both types of individuals choose whether to assimilate.

Table 3.1: Game of Cultural Assimilation

	L-type: assimilate	L-type: do not assimilate
H-type: assimilate	$(U_{H_A}(L_A), U_{L_A}(H_A))$	$(U_{H_A}(L_N), U_{L_N}(H_A))$
H-type: do not assimilate	$(U_{H_N}(L_A), U_{L_A}(H_N))$	$(U_{H_N}(L_N), U_{L_N}(H_N))$

where

$$U_{H_A}(L_A) = 1 + (2n-1)(1-f) - C_H, U_{H_A}(L_N) = 1 + (n-1)(1-f) - C_H \tag{3.4}$$

$$U_{H_N}(L_A) = -1 + (n-1)(1-f), U_{H_N}(L_N) = -1 + (2n-1)(1-f) \tag{3.5}$$

$$U_{L_A}(H_A) = 1 + (2n-1)(1-f) - C_L, U_{L_A}(H_N) = 1 + (n-1)(1-f) - C_L \tag{3.6}$$

$$U_{L_N}(H_A) = -1 + (n-1)(1-f), U_{L_N}(H_N) = -1 + (2n-1)(1-f) \tag{3.7}$$

I now show that H-type individuals assimilate and L-type individuals do not assimilate if and only if $C_H < 2 - n(1-f)$, and $C_L > 2 + n(1-f)$.

Proof. If H-type individuals assimilate and L-type individuals do not assimilate, then we must have (a) $U_{H_N}(L_N) < U_{H_A}(L_N)$, and (b) $U_{L_N}(H_A) > U_{L_A}(H_A)$. Solve the above inequalities, we have (a):

$$-1 + (2n-1)(1-f) < 1 + (n-1)(1-f) - C_H \tag{3.8}$$

and (b):

$$-1 + (n-1)(1-f) > 1 + (2n-1)(1-f) - C_L \tag{3.9}$$

or more simply

$$C_H < 2 - n(1-f); C_L > 2 + n(1-f) \tag{3.10}$$

Note that the above conditions of the equilibrium of cultural assimilation is sufficient and necessary: the opposite case of acculturational homophily—H-type individuals do not assimilate and L-type individuals assimilate—cannot be the equilibrium, as these are not the best responses for both types of individuals.

The above result implies that if the assimilation cost for H-type individuals is low enough, and the assimilation cost for L-type individuals is high enough, we can still observe homophily based on acculturational behaviors—in this study, English-name usage—even if the type of the willingness of assimilation cannot be observed, and the type needs not to be consistent with the acculturational behavior.

3.4.3 Discussions

The above result shows the conditions of the presence of acculturational homophily: H-type individuals tend to acculturate and make friends with other H-type individuals if C_H is small enough; meanwhile, L-type individuals will not acculturate and make friends with other L-type individuals if C_L is large enough.

This result shows some interesting points. First, it provides an explanation for the empirical evidence of "segmented assimilation" among immigrants when social network effects are taken into consideration (Verdier and Zenou, 2015). Specifically, if assuming that there is no terms for friendship in the utility function (which means $u_i(x_i) = x_i - C_{T_i}$), H-type individuals should choose to assimilate when $C_H < 2$, and L-type individuals choose to assimilate when $C_L > 2$. After introducing the term of friendship, H-type individuals become less likely to assimilate, in the sense that the upper bound of C_H for the occurrence of assimilation becomes lower because individuals gain utility from both cultural assimilation and their friendships. This is similar to the theory of "social multipliers" (e.g., Calvó-Armengol and Zenou, 2004).

Moreover, this result points out effects of social networks might lead to different socialization outcomes *among different populations*. Specific empirical examples in the context of the U.S. show that some ethnic groups have relatively higher assimilation costs than European immigrants, and indeed, immigrants from these ethnic groups usually have lower degrees of assimilation even if they are recognized to be "H-type" (Portes and Zhou, 1993; Abramitzky et al., 2014), measured by language proficiency, name usage, education, or other indexes.

Equation 3.10 also implies that we might be able to assume that the assimilation cost has the symmetric structure. Based on this idea, let $C_H = 2 - n(1-f) - \varepsilon$ and $C_L = 2 + n(1-f) + \varepsilon$, where $\varepsilon > 0$. This implies that $C_H + C_L = 4$, or equivalently

$$\frac{C_H + C_L}{2} = 2 \tag{3.11}$$

In the equilibrium, the difference in the utility between H-type and L-type individuals can be described as

$$\Delta u = [1 + (n-1)(1-f) - C_H] - [-1 + (n-1)(1-f)] = 2 - C_H \tag{3.12}$$

or more simply

$$\Delta u = \frac{C_L - C_H}{2} \tag{3.13}$$

This shows that the difference in the utility in the equilibrium is related to the difference in the assimilation cost when the type cannot be observed and individuals decide whether to make assimilation efforts based on the type-specific assimilation cost.

3.5 Data and Empirical Strategies

This section introduces data and empirical strategies. I first analyze two major challenges that are widely seen in studies of homophily. I then discuss the data and methods that tackle these challenges and explain how the extent of homophily will be examined.

3.5.1 Statistical Challenges in Homophily Studies

Generally, there are mainly two statistical challenges in studies that examine the presence of homophily and how the representative characteristics of friendships are shaped. The first challenge is the data issue: many social surveys do not pay much attention to friendships. It is difficult to study homophily in friendships among individuals in developing countries—e.g., in the context of this paper—due to the lack of microdata, but studying homophily can be challenging even in developed countries. There are usually few questions about friendships, or the term "friendship" is not well-defined[7] (Marmaros and Sacerdote, 2006). Some recent surveys (e.g., AddHealth) do ask questions about friendships, but there is still not much information about friends' characteristics. Most studies use administrative data (e.g., registrar records) or communication data (e.g., campus email) because these data provide some individual attributes of both the respondents and their friends. In these papers, demographic homophily has been widely studied (e.g., Marmaros and Sacerdote, 2006; Mayer and Puller, 2008). However, in the setting with not much variation in demographic characteristics (e.g., within an ethnic group), homophily based on non-demographic characteristics is still not well explored. The only related paper with the similar scope is Girard et al. (2015), in which the authors find that personality and background characteristics are robust predictors of friendship characteristics.

The second challenge is methodological: English-name usage might be an endogenous behavior. The reflection problem commonly seen in any so-

[7]This leads to the issue that the role of friendships in the setting of higher education is somewhat debatable in many studies (e.g., Stinebrickner and Stinebrickner, 2006). Specifically, it is not easy to understand how friendships are organized is challenging, and is difficult to even define friendships (Foster, 2006). This further leads to the econometric difficulty when estimating the effect of friends, since even "not-so-close" friends might have disproportionately weak impacts (Mora and Oreopoulous, 2011; Lin and Weinberg, 2014).

cial network research (Manski, 1993) similarly exists in this paper: individual English-name usage can either be the cause or the consequence of close friends' English-name usage. Another concern is that English-name usage might be mis-measured.

I attempt to tackle the above challenges by (a) using a novel online social networking data set, and (b) exploiting the natural experiment on English-name usage. I will introduce data and methods in the remainder of this section.

3.5.2 Data

I tackle the first statistical challenge by using a representative networking data set on students who received undergraduate education in China and, straight afterwards, started graduate education in the U.S. The data set is retrieved from Renren.com, which is widely recognized as the Chinese version of Facebook.

Renren shows users' basic biographical and educational information. It also shows the number of Renren friends and the number of times the personal page is visited by others, which describe the use of social networking. Moreover, Renren has two unique features related to this paper. First, a Renren user has the option to add an English name following the Chinese name as the *suffix*, which measures English-name usage. Second, a user can nominate up to ten *special friends*. Therefore, "close friendships" are well defined because (a) special friends are nominated by users, and can be changed at any time; (b) a Renren user's special friends usually only include friends or significant others, but not parents[8], who are naturally in (but not selected into) individuals' so-

[8]Unlike on Facebook, Chinese parents rarely have Renren accounts. Indeed, most Chinese

cial relationships. Due to data limitation, however, I am unable to determine whether these special friends are in the U.S. or in China. Hence this paper examines homophily among Chinese students but not among Chinese students *in the U.S.*

Some students do not report age and I thus cannot include it as the covariate. However, users must report the starting year of undergraduate and graduate education. Hence, "year since post-secondary education" serves as the proxy for age. It is also highly collinear with year since arrival as I only focus on students who enter the U.S. right after receiving bachelor's degrees. In general, as Renren was founded in 2005 and the sample was retrieved in 2014, 92.2% of students in the sample started their undergraduate programs between 2004 and 2009, implying the high concentration of birth years around 1988.

As Renren mainly provides services for students, it is unpopular among non-students: for a long time Renren was named Xiaonei, which literally means "on-campus network", and the website was open for registration only for students. Similarly, many Renren users seldom use the website after graduation[9]. This implies that English names are mostly added in the period of schooling, and the sample describes English-name usage in *school* friendships. In other words, although Renren does not track students' information after graduation, post-schooling English-name usage (either the student works in the U.S. or returns to China) is also usually not captured in the sample.

Finally, I need to point out two data issues in this paper. First, students

parents are not listed as close friends, which is verifiable as I can observe special friends' profile pictures. I find that almost all students in the sample do not list parents as special friends.

[9]This can be partially tested by checking the date of uploading the latest profile photo, which is publicly available on Renren. Indeed, most users do not upload any new profile photo after graduating from school.

who do not use Renren are clearly not included in the sample. Therefore, I can only examine acculturational homophily conditional on the use of Renren, although if a student does have the Renren account, his networking usage can be controlled. This is probably not a serious issue since Renren is popular among Chinese students and most students have accounts and use Renren when they are in school[10].

Second, there are several types of measurement issues related to English-name usage. A student who does not show the English name online might actually use an English name in real life. There are also ambiguous cases of name identification: it is not rare that non-Anglophone speakers (including Chinese students) adopt non-mainstream or even "weird" names that are not commonly used in Anglophone countries[11]. However, measurement error might still exist even if adding non-mainstream English names into the name dictionary. It is possible that online English-name usage only reflects fashion, or the English name is actually not for personal use (e.g., the "name" is actually the name of the student's idol). In other words, the presence of the English name might not imply English-name usage in reality. In all above cases, English-name usage will be mis-measured.

[10] I conduct a simple test by observing Tsinghua students' Renren accounts. Tsinghua University is one of the few schools in China that release the list of freshmen students online. I find that, indeed, more than 90% of all Tsinghua students register accounts on Renren.

[11] The paper by Edwards and Caballero (2008) describes the general picture of the (weird) adoption and use of the English name by non-Anglophone nationals or immigrants. More related to this paper, there are many non-academic reports and news that discuss the "weird" English names adopted by Chinese people.

3.5.3 The Instrumental Variable (IV) Model

I now turn to the methodological challenge in homophily studies. To analyze econometric issues in this paper, I start with the traditional OLS specification:

$$N_i = \beta_0 + \beta_1 E_i + \beta_2 T_i + \mathbf{X}_i \beta_3 + e_i \tag{3.14}$$

where for student i, N_i is the number of close friends with English-name usage. "Close friends" are defined by the *special friends* nominated by i. E_i is the indicator of English-name usage. T_i is the total number of close friends nominated. T_i is included so that this OLS specification shows whether homophily occurs because of individual behavior rather than the size of the friendship. $\mathbf{X}_i$ is the vector of control variables of individual characteristics, as well as pre- and post-migration characteristics. These control variables will be introduced in the next section. In Equation 3.14, β_1 is the OLS estimate of the impact of individual name usage on the presence of English-name usage in the friendship.

There are econometric issues when estimating the extent of acculturational homophily using OLS. The reflection problem (Manski, 1993) is commonly seen in network research. In this paper, individual English-name usage might be affected by friends' English-name usage. One possible solution to this problem is the random assignment of peers. Some colleges randomly assign freshmen dormitories, which creates natural experiments on peer influence (Marmaros and Sacerdote, 2001; Zimmerman, 2003; Foster, 2006). The random assignment of peers, however, has several drawbacks in homophily studies. First, randomly assigned peers might not actually be friends. Second, in such random assignments, social relationships are pre-determined, but not formed by individuals. Third, randomly assigning peers is not equivalent to randomly assigning in-

dividual behaviors. The random assignment of peers can be useful in studies of peer effects, but as discussed earlier, this is not the only mechanism that accounts for the presence of homophily.

Another methodological issue is measurement error. If student i does not show the English name online but uses an English name in real life, or if i shows an English name online but does not use it in real life, then his English-name usage is mis-measured. Measurement error might also occur if it is difficult to identify whether an English word added by the student is an English name or not.

If the reflection problem is the only issue, then the OLS estimate is upward biased. Taking measurement error into account, the OLS estimate can be either upward or downward biased, and it is unlikely to determine how the OLS estimate is biased. In the following section I will discuss the direction of bias after reporting the main results.

To tackle the problem that E_i is endogenous, in this paper I exploit the language-related natural experiment on English-name usage and employ the instrumental variable (IV) strategy. Specifically, I use the indicator of the pronunciation difficulty of the Chinese name to instrument for English-name usage. Let the pronunciation difficulty dummy variable for i be D_i (where $D_i = 1$ if his Chinese name is "difficult to be pronounced in English", and 0 otherwise). We have the following first-stage regression:

$$\hat{E}_i = \gamma_0 + \gamma_1 D_i + \gamma_2 T_i + \mathbf{X}_i \gamma_3 + \mu_i \tag{3.15}$$

Using $\hat{E}_i$ in the second-stage regression I can thus obtain the IV estimate of $\hat{\beta}_1$ using two-stage least-squares. Note that $\hat{\beta}_1$ captures the contributions of both in-

dividual peer influence and peer selection in shaping acculturational homophily. Due to the cross-sectional nature of the sample, it is unlikely to separate out two types of effects in this paper.

3.5.4 Summary Statistics

In the remainder of this section I will discuss the summary statistics of the sample. I start with individual characteristics in Table 3.2. In the first two columns I describe the full sample in which all students are included, and in the last two columns I report the statistics of the sub-sample that excludes students who list no close friend. In the full sample, 13.3% of all students (approximately 1,000 students) are English-name users on Renren. The mean number of special friends nominated by students is 1.61. On average, students in the sample receive slightly fewer than 10,000 times of visits by other users and have about 600 Renren friends. Approximately 49% of all students are male. In the sub-sample, 15% of all students are English-name users Renren. The average number of close friends, conditional on listing close friends, is around 3.7. Compared with students in the full sample, students in the sub-sample receive more visits from other users and have more Renren friends. Also, there are relatively fewer male students in this sub-sample.

On average, students have started post-secondary education for 8.6 years in both samples. This variable serves as the proxy for age, since I only focus on students who moved to U.S. straight after completing the undergraduate program. The concentration of birth years is roughly around 1988 (the data set was retrieved in 2014). Two samples also share the similar "school-tier composition":

Table 3.2: Summary Statistics: Individual Characteristics

	Full Sample		Sub-Sample	
Independent Variables:	Mean	Std. dev.	Mean	Std. dev.
English-name usage	0.133	(0.340)	0.154	(0.361)
# of special friends	1.610	(2.318)	3.668	(2.165)
# of visits received (in 10,000)	0.935	(1.288)	1.116	(1.418)
# of all friends (in 100)	5.859	(4.815)	6.322	(4.629)
Male	0.489	(0.500)	0.420	(0.494)
Year since post-secondary education†	8.684	(1.826)	8.610	(1.690)
Tier 1 Chinese college	0.205	(0.403)	0.212	(0.409)
Tier 2 Chinese college	0.270	(0.444)	0.267	(0.443)
Tier 3 Chinese college	0.525	(0.499)	0.521	(0.500)
Tier 1 U.S. graduate school	0.140	(0.347)	0.135	(0.342)
Tier 2 U.S. graduate school	0.497	(0.500)	0.523	(0.500)
Tier 3 U.S. graduate school	0.363	(0.481)	0.343	(0.475)
Private U.S. school	0.449	(0.497)	0.447	(0.497)
Observations	7,222		3,171	

The sub-sample excludes students who do not list any special friend on Renren.

†: As discussed earlier, "year since entring college" serves as the proxy for age.

slightly more than 20% of students graduated from tier 1 Chinese colleges, 27% from tier 2 colleges, and 52% from tier 3 colleges. In both samples, approximately 14% of students receive graduate education in tier 1 U.S. schools, half of all students in tier 2 U.S. schools, and 35% in tier 3 U.S. schools. 45% of students enter private schools in the U.S.

Table 3.3 presents summary statistics of the characteristics of geographic areas where students stayed before arriving in the U.S. I categorize China's provinces into five regions[12]. Controlling for students' original areas is impor-

[12]Provinces in *East Coast* include: Jiangsu, Shanghai, Zhejiang, Fujian, Guangdong, Hainan.

tant as regional socioeconomic inequality has long been an issue in China (e.g., Kanbur and Zhang, 1999) and remains crucial even in recent years (e.g., Qin et al., 2016). In both samples, roughly 30% and 45% of students are originally from East Coast and Central North of China. This is not surprising, as both regions are highly populated and are home to most universities in China. Besides, around 7% of students are from Northeast, 10% are from Central South, and 7% are from West. Finally, 29% of students in the sample are originally from China's coastal areas.

Table 3.3: Summary Statistics: Pre-Arrival Geographic Variables (in China)

	Full Sample		Sub-Sample	
Independent Variables:	Mean	Std. dev.	Mean	Std. dev.
Region 1: East Coast	0.313	(0.463)	0.328	(0.469)
Region 2: Central North	0.435	(0.496)	0.444	(0.497)
Region 3: Northeast	0.075	(0.263)	0.079	(0.270)
Region 4: Central South	0.104	(0.305)	0.089	(0.285)
Region 5: West	0.073	(0.260)	0.061	(0.239)
Coastal area	0.287	(0.452)	0.295	(0.456)
GDP per capita (CNY)	97157.570	(20744.100)	97965.600	(19997.520)
Human development index	0.767	(0.056)	0.770	(0.054)
Population (urban area)	1.350e+07	(6.960e+06)	1.370e+07	(6.907e+06)
Area (urban area, sq mi)	324.918	(141.209)	330.364	(138.841)
Observations	7,222		3,171	

The sub-sample excludes students who do not list any special friend on Renren.

I also include local socioeconomic and demographic characteristics as covariates. At the city level, the average GDP per capita (in China's currency) is roughly 100,000, and the human development index (HDI) is around 0.77. Note

Provinces in *Central North* include: Beijing, Tianjin, Shandong, Shanxi, Hebei. Provinces in *Northeast* include: Heilongjiang, Jilin, Liaoning. Provinces in *Central South* include: Henan, Anhui, Jiangxi, Hubei, Hunan. All other provinces are included in *West*.

that these two indexes are much higher than the national average, due to the selection of students in terms of college attendance and migration plans by region. On average, the area of origin has a population of over 10 million, and covers slightly more than 300 square miles.

Table 3.4: Summary Statistics: Post-Arrival Geographic Variables (in the U.S.)

	Full Sample		Sub-Sample	
Independent Variables:	Mean	Std. dev.	Mean	Std. dev.
Population (local)	5.176e+05	(8.768e+05)	5.308e+05	(9.017e+05)
Area (local)	82.908	103.333	83.858	(105.905)
Population (county)	1.856e+06	(2.667e+06)	1.875e+06	(2.717e+06)
% Asian	0.090	(0.064)	0.091	(0.064)
% Chinese	0.036	(0.032)	0.036	(0.032)
% White	0.615	(0.147)	0.614	(0.147)
% Black	0.207	(0.174)	0.208	(0.174)
Median earnings per worker	48647.380	(10409.89)	48770.710	(10543.940)
% Bachelor's or higher degrees	0.470	(0.137)	0.472	(0.139)
Embassy/Consulate of China	0.232	(0.422)	0.239	(0.427)
Avg. # of flights to China per day	0.952	(1.487)	0.967	(1.492)
Observations	7,222		3,171	

The sub-sample excludes students who do not list any special friend on Renren.

Table 3.4 presents descriptive statistics of post-arrival characteristics. On average, the "hosting" local area has a population of over 500 thousand people, and covers about 80 square miles. I also include the county-level population, which is close to 2 million. On average, 9% of local residents have Asian origin, and 3.6% have Chinese origin. Approximately 60% of residents are White and 20% are Black. Again, the average racial makeup is different from the national average due to the geographic distribution of universities in the U.S. On average, the median earnings per worker is close to 50,000 dollars, and 47% of local

residents hold bachelor's or graduate degrees. I finally examine the connection with China. About 23% of students receive education in cities or counties in which China maintains diplomatic missions[13]. On average, there is nearly one direct flight back to China per day in the county of school or nearby counties.

Table 3.5: Summary Statistics: Dependent Variables and IV

	Full Sample		Sub-Sample	
Dependent Variables/IV:	Mean	Std. dev.	Mean	Std. dev.
# of close friends with English-name usage	0.106	(0.371)	0.241	(0.531)
% of close friends with English-name usage	—	—	0.062	0.159
If any close friend uses the English name	0.087	(0.232)	0.199	(0.399)
Pronunciation difficulty indicator	0.422	(0.494)	0.425	(0.494)
Observations	7,222		3,171	

The sub-sample excludes students who do not list any special friend on Renren.

Table 3.5 presents the summary statistics of the dependent variable and the IV. In the full sample, the average number of close friends with English-name usage is about 0.1. In the sub-sample, conditional on listing close friends, this number is 0.24, and on average 6.2% of all close friends are English-name users. In the full sample, 8.7% of students have at least one close friend who uses the English name, and this proportion is nearly 20% in the sub-sample. Finally, the percentage of difficult-to-pronounce names is approximately 42% in both samples. Note that this percentage is very close to that reported in Appendix C, where I investigate external samples of Chinese names that contain non-migrant students.

In Table 3.6, I compare the IV and dependent variables between students who show and do not show English names on Renren. There are 960 students

[13]Unlike most European countries, China does not have any honorary consulate abroad.

Table 3.6: Comparing Students with and without English-Name Usage

	Showing English names	Not showing English names
Full Sample:		
Pronunciation difficulty indicator	0.648 (0.478)	0.388 (0.487)
# of special friends w/ English names	0.401 (0.699)	0.060 (0.262)
Observations	960	6,262
Sub-Sample:		
Pronunciation difficulty indicator	0.640 (0.480)	0.386 (0.487)
# of close friends w/ English names	0.787 (0.809)	0.141 (0.387)
% of close friends w/ English names	0.193 (0.233)	0.038 (0.128)
Observations	489	2,682

The sub-sample excludes students who do not list any special friend on Renren.

Standard deviations are in parentheses.

with English-name usage in the full sample, and 64.8% of them have difficult-to-pronounce Chinese names. On the other hand, only 38.8% of students without English-name usage have difficult-to-pronounce Chinese names. The similar pattern remains in the sub-sample. This table also shows the strong correlation between self English-name usage and English-name usage of close friends, i.e., the presence of acculturational homophily: in both samples, English-name users have more close friends who are also English-name users.

3.5.5 Balancing Tests

In this essay, the indicator of the pronunciation difficulty of the Chinese name serves as the IV for English-name usage. In Table 3.7 I conduct balancing tests to check the systematic difference in individual and geographic characteristics

between two groups of students categorized by the pronunciation difficulty of the name.

In general, Table 3.7 shows no systematic difference between two groups of students. Two groups of students share similar demographic characteristics and have similar levels of Renren usage. In particular, birth years are highly concentrated in both groups and year since starting post-secondary education does not well predict the pronunciation difficulty. There is also no significant difference in the educational background.

Local socioeconomic, cultural, and demographic characteristics might also directly affect the acculturational behavior and acculturational homophily, and there is huge heterogeneity in local characteristics within both China and the U.S. However, I find almost no statistical significance between two groups. In the full sample, I do find that students from Northeast China are slightly more likely to be associated with the pronunciation difficulty, but the population (as shown in Table 3.3 and 3.7) is small and the unbalance can be due to small sample bias. In the full sample I also observe the difference in the percentage of Chinese residents in the (post-arrival) local area, but the difference is subtle. Moreover, in the sub-sample, I do not find the systematic difference in any characteristic between two groups of students. As far as we can observe, the pronunciation difficulty is not associated with any specific age, gender, area (of both origin and destination), and school tier.

Table 3.7: Systematic Differences: Control Variables and the "Pronunciation Difficulty"

	Full Sample			Sub-Sample		
Control Variables:	w/o difficult-to-pronounce names	w/ difficult-to-pronounce names	*p*-value	w/o difficult-to-pronounce names	w/ difficult-to-pronounce names	*p*-value
Individual demographics:						
Male	0.486 (0.500)	0.493 (0.500)	n.s.	0.423 (0.494)	0.417 (0.493)	n.s.
Year since entering college	8.732 (1.822)	8.619 (1.828)	*	8.640 (1.676)	8.581 (1.715)	n.s.
Networking characteristics:						
# of special friends	1.581 (2.292)	1.650 (2.353)	n.s.	3.622 (2.154)	3.730 (2.180)	n.s.
# of visits (in 10,000)	0.933 (1.315)	0.937 (1.250)	n.s.	1.095 (1.428)	1.144 (1.403)	n.s.
# of friends (in 100)	5.820 (4.544)	5.913 (5.163)	n.s.	6.323 (5.247)	6.321 (3.635)	n.s.
School information:						
Tier 1 Chinese college	0.203 (0.402)	0.207 (0.405)	n.s.	0.205 (0.404)	0.222 (0.416)	n.s.
Tier 2 Chinese college	0.271 (0.445)	0.269 (0.443)	n.s.	0.268 (0.443)	0.265 (0.442)	n.s.
Tier 3 Chinese college	0.526 (0.499)	0.524 (0.500)	n.s.	0.527 (0.499)	0.513 (0.500)	n.s.
Tier 1 U.S. graduate school	0.138 (0.344)	0.144 (0.351)	n.s.	0.131 (0.338)	0.140 (0.347)	n.s.
Tier 2 U.S. graduate school	0.498 (0.500)	0.495 (0.500)	n.s.	0.518 (0.500)	0.528 (0.499)	n.s.
Tier 3 U.S. graduate school	0.364 (0.481)	0.361 (0.480)	n.s.	0.351 (0.477)	0.332 (0.471)	n.s.
Private school in the U.S.	0.445 (0.497)	0.454 (0.498)	n.s.	0.441 (0.497)	0.454 (0.498)	n.s.
Pre-arrival geo. variables:						
Region 1: East Coast	0.307 (0.461)	0.318 (0.466)	n.s.	0.328 (0.470)	0.327 (0.470)	n.s.
Region 2: Central North	0.432 (0.495)	0.438 (0.496)	n.s.	0.440 (0.497)	0.448 (0.498)	n.s.
Region 3: Northeast	0.069 (0.254)	0.082 (0.275)	*	0.074 (0.262)	0.086 (0.280)	n.s.
Region 4: Central South	0.104 (0.305)	0.104 (0.305)	n.s.	0.095 (0.294)	0.080 (0.271)	n.s.
Region 5: West	0.071 (0.258)	0.074 (0.263)	n.s.	0.062 (0.241)	0.058 (0.234)	n.s.
Coastal city	0.290 (0.454)	0.283 (0.451)	n.s.	0.301 (0.459)	0.287 (0.452)	n.s.
Log GDP per capita (CNY)	11.459 (0.263)	11.447 (0.273)	n.s.	11.463 (0.259)	11.465 (0.260)	n.s.
Human development index	0.768 (0.056)	0.766 (0.055)	n.s.	0.770 (0.055)	0.771 (0.054)	n.s.
Log population (urban)	16.243 (0.679)	16.221 (0.675)	n.s.	16.244 (0.689)	16.263 (0.645)	n.s.
Area (urban, sq mi)	327.315 (141.561)	321.637 (140.685)	n.s.	328.555 (139.706)	332.808 (137.679)	n.s.
Post-arrival geo. variables:						
Log population (local)	12.206 (1.332)	12.229 (1.374)	n.s.	12.204 (1.336)	12.251 (1.392)	n.s.
Area (local, sq mi)	82.843 (103.393)	82.996 (103.267)	n.s.	83.010 (104.636)	85.002 (107.625)	n.s.
Log population (county)	13.626 (1.259)	13.616 (1.236)	n.s.	13.604 (1.623)	13.629 (1.252)	n.s.
% Asian residents	0.089 (0.063)	0.092 (0.066)	n.s.	0.089 (0.063)	0.093 (0.066)	n.s.
% Chinese residents	0.035 (0.032)	0.037 (0.033)	*	0.035 (0.032)	0.037 (0.033)	n.s.
% White residents	0.618 (0.147)	0.611 (0.147)	n.s.	0.617 (0.148)	0.611 (0.145)	n.s.
% Black residents	0.206 (0.174)	0.209 (0.175)	n.s.	0.208 (0.174)	0.206 (0.173)	n.s.
Log median earnings	10.771 (0.207)	10.772 (0.198)	n.s.	10.773 (0.212)	10.772 (0.195)	n.s.
% Bachelor's or higher	0.470 (0.137)	0.471 (0.137)	n.s.	0.472 (0.140)	0.473 (0.139)	n.s.
Embassy/Consulate of China	0.231 (0.422)	0.233 (0.423)	n.s.	0.232 (0.426)	0.242 (0.429)	n.s.
Avg. # of flights to China	0.951 (1.502)	0.952 (1.468)	n.s.	0.962 (1.497)	0.974 (1.484)	n.s.
Observations	4,173	3,049		1,822	1,349	

Standard deviations are in parentheses. Unpaired *t* tests are employed. n.s.: $p \geq .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

3.6 Empirical Analysis: Acculturational Homophily

In this section I will examine the extent of homophily based on English-name usage. I first present the first-stage regressions, and then report the empirical results of this paper. I measure the extent of homophily by the number of close friends with English-name usage, and examine whether self English-name usage affects this extent. Subsequently I focus on the econometric issue of this paper again and discuss the direction of OLS bias. I conclude this section by examining heterogeneous effects in some subpopulations, and in Appendix E I provide other additional tests for the results.

3.6.1 The First-Stage Relationship

I first report the first-stage regressions in Table 3.8. The pronunciation difficulty indicator serves as a valid IV for English-name usage only if two variables are closely correlated. In Column 1 I run the first-stage regression in the full sample and add only individual characteristics school tier fixed effects as covariates. All else being equal, a student whose Chinese name is difficult to be pronounced is 12.1% more likely to show his English name, compared with a student without the difficult-to-pronounce name. The pronunciation difficulty indicator well predicts English-name usage, and the F-statistic is comfortably greater than 10. Column 2 repeats the exercise in the sub-sample in which students list at least one close friend. Similarly, in this sub-sample I find that a student with the difficult-to-pronounce Chinese name is more likely to be the English-name user on Renren.

Table 3.8: First-Stage Regressions

	English-name usage indicator					
	(1)	(2)	(3)	(4)	(5)	(6)
Pronunciation difficulty	0.121***	0.132***	0.121***	0.133***	0.120***	0.133***
	(0.008)	(0.013)	(0.008)	(0.013)	(0.008)	(0.013)
Individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School covariates	Yes	Yes	Yes	Yes	Yes	Yes
Pre-arrival characteristics	No	No	Yes	Yes	Yes	Yes
Post-arrival characteristics	No	No	No	No	Yes	Yes
Sample	Full	Sub	Full	Sub	Full	Sub
First-stage F-statistic	238.27	111.38	237.69	113.76	236.21	112.29
Observations	7,222	3,171	7,222	3,171	7,222	3,171

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In the remainder of this table I add pre- and post-arrival geographic characteristics into the model. In general, I find the quantitatively similar first-stage relationship after including pre-arrival (Column 3 and 4) and additionally post-arrival variables (Column 5 and 6). Hence, the pronunciation difficulty indicator is a valid IV in the sense that it well predicts English-name usage. Note the IV estimate based on this strategy is the *local* average effect of English-name usage (i.e., LATE).

3.6.2 Main Results

I now turn to the main empirical results of this paper. In Column 1, Table 3.9, I focus on the full sample and run a baseline OLS regression. I estimate the effect of self English-name usage on the presence of acculturational homophily using OLS, and add individual and school characteristics as covariates. The result

shows the positive correlation between self English-name usage and English-name usage of close friends. In Column 2 I rerun the regression but using the difficulty of pronouncing the Chinese name as the IV. Again, I observe the presence of acculturational homophily based on English-name usage, and the OLS estimated is downward biased.

Table 3.9: Homophily based on English-Name Usage: OLS and IV Models

	Number of close friends with English-name usage					
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	IV	OLS	IV	OLS	IV
English-name usage	0.291***	0.434***	0.290***	0.437***	0.584***	0.917***
	(0.011)	(0.064)	(0.011)	(0.064)	(0.023)	(0.127)
# of close friends	0.063***	0.061***	0.062***	0.061***	0.055***	0.051***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
Individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School covariates	Yes	Yes	Yes	Yes	Yes	Yes
Pre-arrival characteristics	No	No	Yes	Yes	Yes	Yes
Post-arrival characteristics	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	Full	Sub	Sub
R^2	0.258	—	0.259	—	0.262	—
Observations	7,222	7,222	7,222	7,222	3,171	3,171

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In Column 3 and 4 I control for geographic characteristics and repeat the exercise. I find the similar magnitude of the effect of self English-name usage, and the OLS estimate appears to be downward biased. To exclude students who have no close friend listed online, in Column 5 and 6 I focus only on the sub-sample and examine the extent of homophily again, and find that self English-name usage leads to nearly one more special friend who also uses the English name, controlling for all other characteristics. This magnitude is fairly large

relative to the size of the friendship: in the sub-sample, the average number of close friends is about 3.7 (see Table 3.2). Note that in all regressions I include the total number of close friends, which means that the effect of individual behavior on the presence of homophily is not through the size of the friendship overall.

Table 3.9 indicates that individuals actively shape the representative characteristic—English-name usage—of friendships. The results are based on the traditional OLS as well as the IV estimate that fixes the issue of reversal causality, although both peer influence and peer selection can explain how individual behavior leads to group identity.

3.6.3 Discussions and Additional Tests

I conclude this section by discussing the bias issue of OLS regressions and then conducting several additional tests. I first revisit the endogeneity problem: as analyzed earlier, in theory it is unlikely to predict the direction of bias when estimating the effect of self English-name usage using OLS. Table 3.9 shows that the OLS estimate is downward biased in every model. This indicates that the reflection problem is not the dominant source of endogeneity, because otherwise the OLS estimate should be upward biased.

In this paper, it is likely that measurement error that creates attenuate bias tends to be more important. If self English-name usage is indeed positively correlated with the number of close friends with English-name usage, then the OLS estimate is downward biased if (a) some students who show English names online are actually not English-name users in real life, or (b) some students who do not show English names online are actually English-name users in real life,

assuming that the reflection problem is fixed.

Table 3.10: Measurement Error and the Direction of Bias

	Number of close friends with English-name usage					
	(1)	(2)	(3)	(4)	(5)	(6)
Non-name suffix	−0.064***	−0.065***		−0.126***	−0.127***	
	(0.014)	(0.014)		(0.029)	(0.029)	
Name/non-name suffix			0.156***			0.318***
			(0.010)			(0.021)
# of close friends	0.066***	0.066***	0.063***	0.063***	0.063***	0.058***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)
Individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School covariates	Yes	Yes	Yes	Yes	Yes	Yes
Pre-arrival characteristics	No	Yes	Yes	No	Yes	Yes
Post-arrival characteristics	No	Yes	Yes	No	Yes	Yes
Sample	Full	Full	Full	Sub	Sub	Sub
R^2	0.192	0.194	0.220	0.112	0.118	0.175
Observations	7,222	7,222	7,222	3,171	3,171	3,171

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

It is difficult to analyze (b) because the econometrician cannot observe English-name usage in real life if actual English-name users even do not present names on Renren. In this table, however, I attempt to analyze (a) by examining how the number of close friends with English-name usage is affected by the p-resence of suffixes that are considered to be very unlikely to be names used in real life. Such suffixes include: (a) *last* names of celebrities (e.g., "Einstein"); (b) numbers or letter abbreviations that do not contain any consonant (e.g., "641", "XXY"); (c) academic/professional concepts that are not found to be used by native speakers of Western languages as names (e.g., "Covariance"). These words are most unlikely to be students' English names in real life.

Table 3.10 shows that in both the full sample and the sub-sample, using the non-name suffix does not improve the number of close friends with English-name usage; in contrast, it even leads to fewer friends who are English-name users. In Column 3 and 6 I construct a new "suffix" variable that combines both identified names and non-name suffixes. Using this dummy as the covariate, I find that the effect of English-*suffix* usage is significantly smaller than the effect of English-*name* usage shown in Table 3.9. This similarly implies the negative relationship between the presence of the non-name suffix and the number of close friends who are English-name users.

Results shown in Table 3.10 explain how measurement error might determine the direction of bias of OLS estimates. As reported in Table 3.10, showing an English word that is clearly not a name is negatively correlated with the number of close friends with English-name usage. The more ambiguous cases involve English words that are not names used by students in real life, but resemble actual names and are thus incorrectly identified as names, such as *first* names of celebrities (e.g., "Albert" of Albert Einstein), and words that are academic/professional concepts but have also been used as names (e.g., "Allegro" and "Apriori"). The effect of actual English-name usage is underestimated using OLS if such non-name words are misidentified as names. As it is difficult to exclude all ambiguous cases, in this paper measurement error in English-name usage might play a crucial role and make OLS estimates generally downward biased.

I conclude this section by conducting several additional tests for the heterogeneous effect. Other additional tests are presented in Appendix E of the dissertation. In Table 3.11, I focus on the sub-sample in which only students who

list nonzero close friends are included. In the sub-sample, I further split the population to examine the heterogeneous effect of self English-name usage. In Column 1 I focus on students from tier 1 and 2 Chinese colleges, and in Column 2 I focus on students in tier 1 and 2 U.S. schools. In both columns I find that the IV estimate of the effect of self English-name usage is very close to the average effect reported earlier. This implies that acculturational homophily appears to be universal among students regardless of the tier of the school attended.

Table 3.11: Additional Tests: Heterogeneous Effects (IV Regressions)

	Number of close friends with English-name usage					
	(1)	(2)	(3)	(4)	(5)	(6)
English-name usage	0.916***	0.920***	0.633**	1.124***	0.645**	1.059***
	(0.180)	(0.142)	(0.188)	(0.184)	(0.197)	(0.168)
# of close friends	0.038***	0.051***	0.040***	0.059***	0.048***	0.054***
	(0.006)	(0.005)	(0.007)	(0.007)	(0.006)	(0.006)
Subpopulation:	Tier 1 & 2		Entering college		Gender	
	college	grad school	≤ 2006	> 2006	male	female
Individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
School covariates	Yes	Yes	Yes	Yes	Yes	Yes
Pre-arrival characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Post-arrival characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,519	2,084	1,601	1,570	1,332	1,839

Only students who list nonzero close friends ("sub-sample") are included.

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In Column 3 and 4 I split the sample by the year of starting post-secondary education. This variable is also closely related to age and year since arrival. In Column 3 I focus on students entering college in or before 2006. While the effect is still significantly positive, its magnitude is relatively small. In fact, it is only about two thirds of the average effect presented in Table 3.9. In contrast,

the magnitude of the effect of English-name usage is much larger among students entering the college after 2006. This implies the huge age heterogeneity in the presence of acculturational homophily: although homophily exists in both groups, the effect of self English-name usage is much larger among younger students, even if older students moved to the U.S. earlier.

Finally, I study gender and acculturational homophily in Column 5 and 6. In both columns I find that self English-name usage leads to more friends who also use English names. However, the extent of homophily is much greater among female English-name users. In other words, female students are more likely to develop the representative characteristics of friendships by individual behaviors in the context of this paper.

I conduct several other tests to check the robustness of the results in Appendix D. Using school fixed effects, I reexamine the effect of self English-name usage on the number of close friends with English-name usage, and observe the similar effect size. I also introduce other measures of the extent of homophily, and find the similar qualitative pattern.

3.7 Conclusion

English-name usage is a typical acculturational behavior among immigrants from non-Anglophone countries. Researchers have long observed that English-name usage benefits immigrants—especially those from developing countries—in the labor market and the society of the host country. Moreover, it is expected that positive peer effects can be generated in social relationships defined by name usage.

This paper examines the extent of homophily based on English-name usage, i.e., acculturational homophily, among Chinese students. I use data from Renren.com, a Facebook-type website based in China. The sample consists of students who receive undergraduate education in China and graduate education in the U.S. Renren users have the option to add English names, which measures English-name usage. I can also define "close friends" by users' self-nominated *special friends*. This paper examines the effect of self English-name usage on the number of close friends with English-name usage.

It is methodologically challenging to study how individual's English-name usage is causally related to the number of friends with English-name usage because of the endogeneity problem. Individual English-name usage might be reversely affected by friends' English-name usage. There are also issues that English-name usage can be mis-measured using online data. I exploit a natural experiment to tackle these problems: all else being equal, a Chinese student with the difficult-to-pronounce Chinese name is more likely to use an English name. The pronunciation difficulty is with respect to native speakers of English. I conduct the balancing tests and find that difficult-to-pronounce names are nearly randomly "assigned" in the sample.

The empirical findings of this paper can be summarized as follows. Both the OLS and IV model show that on average, an English-name user has more close friends who are also English-name users, and acculturational homophily based on English-name usage does exist. The OLS estimate is downward biased, and measurement error is the main source of endogeneity. On average, a student who shows the English name online has nearly one more close friend who is an English-name user, conditional on that the student lists nonzero close friends.

The extent of acculturational homophily is large, as the average number of nominated close friends in the sample is approximately 3.7. As I control for the total number of friends and close friends in all models, the effect of self English-name usage is not through the size of the friendship overall.

This paper adds to the literature of development economics and network economics by providing new evidence of acculturational homophily in friendships of migrants from developing countries. It also presents some results of the causal association between individual behavior and group identity. A promising avenue for future research is to decompose the influence of individual behavior on the presence of homophily using longitudinal or experimental data.

CHAPTER 4

A STUDY OF THE EFFECT OF SOCIAL NETWORKS AMONG IMMIGRANTS: ETHNIC SOCIAL NETWORKS AND IMMIGRATION OF HIGH-SKILLED PROFESSIONALS

4.1 Abstract of the Study

This paper investigates effects of ethnic social networks on migration outcomes of French football players in the England. The network size is measured by the number of French teammates. I find that the achievement of the France national team predicts the influx of French players, but not player quality; based on this, I design an instrumental variable (IV) identification strategy. I observe no significant network effect on the outcome of staying in the same football team, but ethnic networks do help French players stay in England. I find heterogeneous network effects, and ethnic networks appear not to always help those most in need.

4.2 Introduction

Immigration has been an important topic in economics since Chiswicks early research (1978, 1980). Economists generally follow two directions in immigration research. The first direction is to examine immigrants' social and economic outcomes in the host society (e.g., Borjas, 1987, 1995, 1998; Lubotsky, 2005; Damm, 2014). The second direction is to examine the impact of immigration on native workers (e.g., Greenwood and McDowell, 1986; Card, 1990; Altonji and Card,

1991; Edo, 2002; Ottaviano, 2004; Ottaviano and Peri, 2006; Mocetti and Porello, 2010). Following the first direction, I study an empirical question about immigration of high-skilled professionals: do ethnic networks in English *football*[1] teams affect French football players' migration outcomes?

There is a vast migration literature on player movement in the sports labor market. Many sports economic studies examine "market effects" on labor mobility. For instance, local tax rates are found to affect internal and international migration of athletes (e.g., Kopkin, 2012; Kleven et al., 2013). Labor market policy shocks, such as the inception of the "free agency"[2] (Depken II, 2002) and the "Bosman Ruling"[3] (Frick, 2009) can also affect player movement. In this paper, I provide an alternative perspective by focusing on social networks that have "non-market" effects (Glaeser and Schienkman, 2002). I put special focus on French players in England because: (a) due to language barriers, it is harder for French players to assimilate into England than German, Dutch, or Scandinavian players; (b) only France "provides" an adequately large sample for our study.

Though little is known about network effects on migration of athletes, the relationship between social networks and immigration is not a new topic in economics. Most prior research does find network effects on low-skilled immigrants (e.g., Munshi, 2003; Damm, 2009). However, it is somewhat difficult to examine the overall network effect among all high-skilled immigrants due to the huge heterogeneity across industries and professions. A possible strategy is to study one single occupation (e.g., Moser et al., 2014). This paper shed-

[1] In this paper, *football* is the popular sport commonly played in Europe. This is different from American football, and in the United States and Canada, the sport is usually called *soccer*.

[2] Mainly seen in the U.S., players are free to sign with any team under certain contract status according to the rule of the "free agency".

[3] Made in 1996, the "Bosman Ruling" banned limitations on purchasing foreign European Union (EU) football players for all football teams in the EU.

s light on network effects among immigrants in the upper tail of the skill and earnings distribution by focusing on migrant football players. The advantage of using sports data is that athletic teams are usually smaller than regular firms, and players have stronger bargaining power than many other types of workers. Teams also have to value and respect the "team culture", including ethnic networks, before any trade involving players.

This paper relies on the instrumental variable (IV) strategy to tackle the typical challenge of the endogenous network (e.g., Manski, 1993; 2000). More specifically, I use a constructed achievement variable that measures the recent achievement of the France national team upon arrival of each French teammate to instrument for the network size. This is based on our argument that the achievement of the France national team is not only the signal for the quality of top French players, but is also positively correlated with the influx of French players with *all* levels of skills. Given that most French players are not good enough to play for their national team, the achievement variable predicts the number of arriving French players, but not their quality. I attempt to ensure the validity of the IV along two dimensions: (a) I regress individual performance in the English league on France's achievement and find no significant effect; and (b) I include a large variety of demographic (e.g., age, year of prior stay), athletic (e.g., position, performance) characteristics and year fixed effects to exclude other channels of the effect. In particular, I control for athletic performance of both the player himself and his French teammates. That said, this paper is not about "peer effects on achievement" and cannot draw any conclusion on the causes and consequences of French player's athletic performance.

The empirical findings of this paper can be summarized as follows. Using

player-year data, I find no clear evidence that the effect of French ethnic social networks on the outcome of staying in the same English football team, but ethnic networks do help French players stay in England. I also find heterogeneous network effects on migration outcomes, and ethnic networks do not always help those most in need. For example, I find no significant network effect on migration outcomes of veteran players or players with few league appearances.

The rest of the paper is organized as follows. Section 4.3 describes the background. Section 4.4 proposes a simple theoretical framework that guides the empirical analysis and points out discusses the econometric specification. Section 4.5 introduces data and methods. Section 4.6 reports the results of network effects on migration outcomes. Section 4.7 concludes.

4.3 Background

In this section, I introduce the institutional background of this paper. I first briefly review prior research on social networks and immigrants' economic and social outcomes. I then discuss the demand and supply side of the sports labor market in this paper: the English Premier League and the French immigrant players.

4.3.1 Social Networks and Immigration

In recent decades it has been increasingly recognized that the ethnic social network plays an important role in shaping immigrants' social and labor market outcomes. Before any empirical analysis, we first need to understand the e-

conomics of ethnic social networks. It has been observed that different racial and ethnic groups tend to reside and work in spatially segregated areas (e.g., Schelling, 1969, Stark, 1991; Borjas, 1999; Gross and Schmitt, 2003). In particular, immigrants living or working in geographically close areas form ethnic social networks, which in turn influence their social and economic outcomes. Are social network effects positive, or negative? This is generally an empirical question and there are two main theoretical hypotheses. The first hypothesis is that immigrants benefit from ethnic networks. The second hypothesis is that immigrants get hurt by living or working within networks.

Both hypotheses can be theoretically correct. Edin et al. (2003) find that labor market outcomes are improved when less-skilled immigrants in Sweden live in ethnic enclaves. Damm (2009) finds similar results using Danish data. Munshi (2003) studies Mexican migrants in the U.S. and finds that immigrants receive supports from compatriots and thus become more likely to be employed and get better jobs. Similar results are shown in other topics, such as academic achievement (Friesen and Krauth, 2010) and job turnover patterns (Hellerstein et al., 2014). Positive network effects are usually generated by knowledge and information spillovers (e.g., Borgatti and Cross, 2003; Munshi, 2004; Conley and Udry, 2010) or collaborative behaviors in the process of cultural assimilation (e.g., Hoxby, 2000; Verdier and Zenou, 2015).

On the other hand, ethnic concentrations are usually associated with discrimination or negative attitudes towards immigrants (Dustmann and Preston, 2001), and one major explanation for this is that immigrants living in ethnic enclaves are more likely to keep their original identities (Battu and Zenou, 2010), and minority identities are further related to discrimination (e.g., Rubinstein

and Brenner, 2014). In addition, network effects can be negative because immigrants living in ethnic enclaves are less willing or needed to acquire certain job skills, such as language ability (e.g., Lazear, 2007). Finally, ethnic enclaves are sometimes related to many social problems. Cutler and Glaeser (1997) show that minority residents living in less segregated areas have better educational and labor market outcomes than those residing in more segregated areas. Patacchini and Zenou (2012) find the positive correlation between the crime rate and the density of the black immigrant population, although Bell and Machin (2013) find exactly opposite empirical results.

While there is a vast literature on the network effect on low-skilled immigrants, it is much more difficult to find a general conclusion on the overall network effect among high-skilled immigrants due to the heterogeneity across industries and professions, although economists have successfully analyzed the social network effect on high-skilled immigrants in specific fields, such as academia. Moser et al. (2014) present positive peer effects on scientific achievement among Jewish migrant chemists; similar effects on the number of publications are found among migrant mathematicians from the Soviet Union (Borjas and Doran, 2012) and China (Borjas et al., 2015), although such positive effects on quantity do not always leads to positive effects on quality (Freeman and Huang, 2015). That said, academia is a very special field and it is substantially different from many other types of fields requiring high professional skills—the outcome in academia is usually measured based on individuals, while for many other types of workers, including football players, the outcomes are mainly measured based on the achievement of the entire firm/team.

4.3.2 English Premier League: The Demand Side

I now focus on the background of the labor market and introduce the demand and supply side, namely the English football league and French football players. In the early 1990s, the then "English First League" was promoted to the Premier League. After that, the English Premier League soon became one of the most successful leagues in Europe. Since the victory of Manchester United in the European Champions League (the most prestigious club-level football tournament) in 1999, there have been four winners and eight finalists from England in this tournament. However, in sharp contrast to the prestige of its league, the England national team is widely considered to be underachieving. After 1990, England has reached semifinal only in the 1996 European Football Championship, and even did not qualify for the 1994 World Cup and the 2008 European Football Championship. Because only English players are eligible to play for the England national team, this reflects that England probably cannot produce adequate high-profile players.

One solution for English teams is to purchase foreign players. After the "Bosman Ruling" came into effect, English football teams have no restriction on buying players from the EU. In Figure 4.1 I indeed observe a sharp increase in the share of foreign players[4] in England after 1995, the last year before the "Bosman Ruling" was made. In particular, after 2000 the share of foreigners is around 60%, and the trend has been stable since then.

[4]Due to historical reasons, England, Scotland, Wales and Northern Ireland have individual football associations, as well as individual "national" teams eligible for national team tournaments. Scottish, Welsh and Northern Irish players are sometimes considered to be "foreigners" in the English Premier League. However, unlike those originally from other EU states, there was no limitation on the number of football players purchased from other parts of the U.K. even before the "Bosman Ruling" was made.

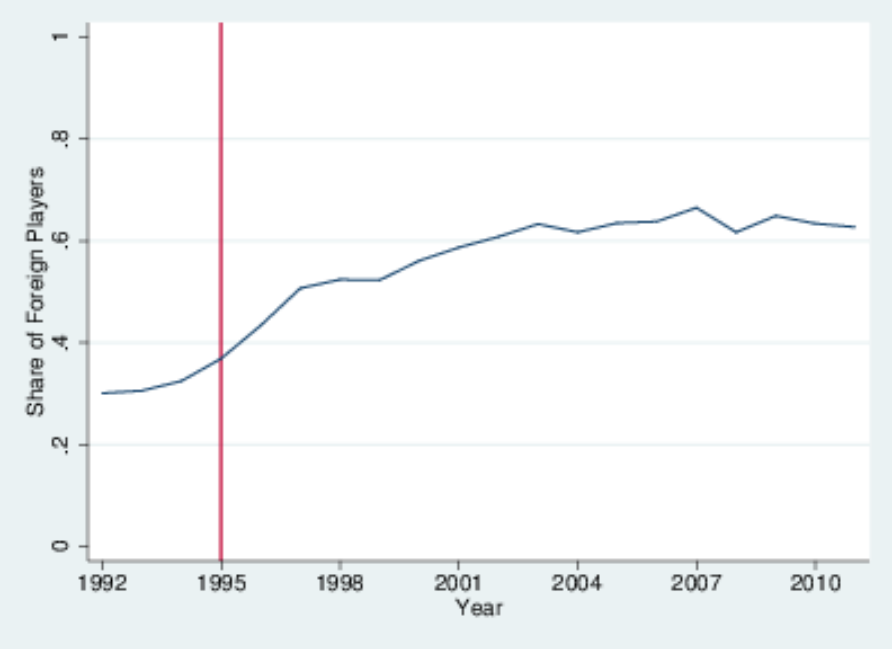


Figure 4.1: The ratio of international football players (defined by nationality) in the English Premier League.

4.3.3 French Players: The Supply Side

Besides the demand of foreign players in the English football market, the supply side has also driven the influx of French players in England. The France national team is one of the most successful teams in football tournaments for national teams in past two decades, having been the winner of 1998 World Cup, 2000 European Football Championship and the runner-up of 2006 World Cup, as well as the two-time winner of the Confederations Cup. France has also not missed any major tournament since 1994.

A key factor that drives the influx of French players is that the French league is relatively less prestigious in Europe. For example, there have been only two French teams reaching the final of the European Champions League (Marseille in 1993 and Monaco in 2004). Indeed, as a comparison, I do not observe the

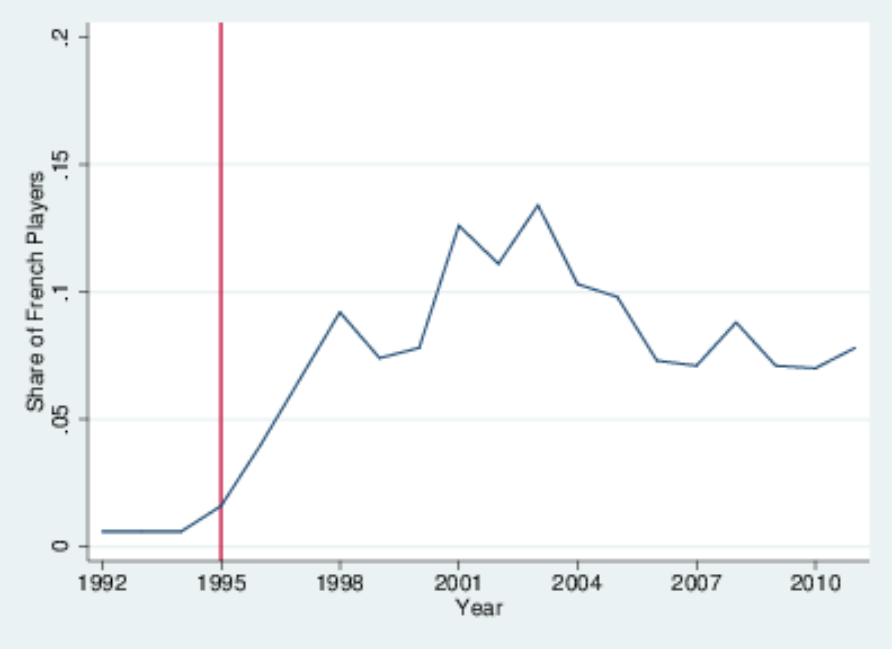


Figure 4.2: The ratio of French players in the English Premier League.

mass influx of German, Italian and Spanish players (though they are all strong national teams comparable with France), and one crucial reason is that football leagues in these countries are similarly prestigious as the English league, hence native players need not to migrate for playing in better leagues. Figure 4.2 shows the share of French players in the English Premier League. I again observe a huge increase in the share of French players in the English Premier Lague after the "Bosman Ruling" was made, and the share of French players is generally around 10%, indicating that the French ethnic group is an important component in the English football market.

4.4 Theoretical Considerations

In this section, I present a simple theoretical framework to consider the identification of network effects. I mainly follow models of Edin et al. (2003) and Munshi (2003) to study immigrants' utility and their migration behaviors.

Suppose an immigrant player has the original utility u prior to arrival. After moving to England, assume that he obtains utility U, and he stays if $U > u$; otherwise, he leaves. For simplicity, assume the linear probability of stay:

$$Pr(X_i = 1) = \beta n_i + w_i + \varepsilon_i + L_E \tag{4.1}$$

where $X_i = 1$ means this i-th immigrant player stays in the host country, $X_i = 0$ otherwise. n_i is the network size, i.e., the number of compatriot teammates. w_i is the wage. ε_i is a variable of non-pecuniary characteristics, i.e., individual information (such as age and previous stay), local social changes, and other factors that may be correlated with $Pr(X_i = 1)$. L_E is a constant describing the prestige of the English Premier League. I want to identify the network effect β.

I make a few assumptions here. The utility in the host country U_i follows some distribution, and $Pr(X_i = 1) = f(U_i)$, where f is a strictly increasing function, i.e., the higher the utility a player gains, the more likely he will remain in the host country. In addition, following a typical principal-agent model (Holmstrom, 1979) I suppose the wage is related to outputs in the English Premier League (which is the league appearances of this player), denoted as m_i, and $w_i = \gamma m_i + \eta_i$, where $\gamma > 0$. This assumption means that the number of matches played positively affects the income. To estimate the network effect, I need to

check

$$\frac{\partial Pr(X_i = 1)}{\partial n_i} = \beta + \frac{\partial(\gamma m_i + \eta_i + \varepsilon_i)}{\partial n_i} = \beta + \gamma\frac{\partial m_i}{\partial n_i} + \frac{\partial \varepsilon_i}{\partial n_i} \tag{4.2}$$

If $\partial m_i/\partial n_i = 0$ and $\partial \varepsilon_i/\partial n_i = 0$, I can describe each player's migration decision by solely looking at the sign of β. The player will stay in the host country if

$$U_i = f^{-1}(\beta n_i + w_i + \varepsilon_i + L_E) > u_i \tag{4.3}$$

Fixing other parameters, the player will stay if $n_i > n_i^0$ when $\beta > 0$, where the threshold size n_i^0 is such that $U_i(n_i^0) = f^{-1}(\beta n_i^0 + w_i + \varepsilon_i + L_E) = u_i$. Similarly, when $\beta < 0$ the player will stay if $n_i < n_i^0$. Whether this player needs a sufficiently big or small social network depends on the sign of β, which can be estimated by an OLS or logit regression. Similarly, I can turn to focus on a player's migration behaviors in his entire career. If assuming time-invariant parameters, his stay in England follows a binomial distribution, where repeat migration (Kirdar, 2009; Constant and Zimmermann, 2012) is allowed. The duration of stay D_i is thus also a linear function of n_i, and I can run a simple regression of the number of years he stays in England on the network size to estimate the network effect β.

However, it is possible that $\partial m_i/\partial n_i \neq 0$ and $\partial \varepsilon_i/\partial n_i \neq 0$. The first inequality implies that a player's league appearances might be correlated with the network size. For example, a French player staying in a team with a larger French network might have fewer problems of communication or playing styles, and thus he is more likely to be fielded on the pitch and has higher outputs. On the contrary, the team manager might try to reduce the number of foreign players based on the team tradition or even his personal preferences (the "domestic bias"). ε_i might also be related to the network size. For example, if England starts to create French-speaking schools, the network size can be larger because

it is easier for French players to bring children to England. Hence, the network size might be endogenous. Still, if knowing the sign of $(\beta + \gamma\partial m_i/\partial n_i + \partial\varepsilon_i/\partial n_i)$ I can similarly find the threshold network size n_i^0. However, the endogeneity problem complicates the analysis and makes OLS or logit estimates biased. In fact, as analyzed above, it is even hard to determine whether the OLS estimate of the network effect is upward or downward biased.

Our strategy to solve this problem is to find an instrumental variable for the network size n_i. In practice, this IV must be independent of m_i and ε_i, but correlated with n_i; also, I impose the exclusion restriction that the IV affects the migration decision only through its association with the network size. In the empirical analysis, I use the achievement of the France national team as the instrumental variable for the network size. This will be discussed in detail in the next section.

4.5 Data and Empirical Strategies

4.5.1 Data

In this paper, I use a player-year data set that contains 141 French players in the English Premier League from 1995 to 2011. I collect a player's individual characteristics in each athletic year (or the so-called *season*) spent in England, including his age, year he has stayed in England ("previous stay"), skin, league appearances, and the playing position. Moreover, if this player just arrives in England, I quantify and record the achievement of the France national team at

that year[5], which will be discussed in detail later. In addition, I collect a number of team-level variables, including the rank, points, and whether the head coach of the team is French. I use the number of French teammates (i.e., the size of the ethnic social network) to measure the network strength, and also record these French teammates' league appearances.

The descriptive statistics of player-year data are shown in Table 4.1. Panel A presents individual characteristics. Players are heterogeneous in league appearances. The total number of matches in one year is 38, and in Table 4.1 I see that there are French players who play in every match, as well as players who rarely play. Players' ages are concentrated in a small interval, in which football players normally enjoy the best time of the career. The average length of previous stay in England is 2.249 years. Black players comprise nearly 60% of all player-year records in data. Finally, only a small fraction of French players have played for France in the latest national team tournament.

In Panel B I present network characteristics. The average network size is slightly more than 2, and there is indeed some heterogeneity in the network size: the maximum number of network members is 6, but there are also French players with no ethnic networks in their teams. The average league appearances in the ethnic network, conditional on non-zero French teammates, is approximately 17 matches. The average league rank is 8.37 (with 20 teams in the league) and the average team points is close to 60 points[6].

[5]For each player-year observation I consider his latest arrival because theoretically, similar to other immigrants (e.g., Constant and Massey, 2003; Kirdar, 2009), French football players are likely to migrate to England and then leave back and forth. Nevertheless, it is worth noting that the proportion of players with experiences of return migration is fairly low. Actually, approximately 95% French players who have ever played in England after 1992 have only one single period of career spent in England.

[6]Generally, an English football team needs at least 85 points to secure the league champion title, and 60 points to secure top 6, although these largely vary by year.

Table 4.1: Descriptive Statistics: Player-Year Data

	Mean	Std.	Max	Min
Panel A: Individual Characteristics†				
Age	26.398	4.205	37	13
League appearances	20.659	11.160	38	1
Previous stay in England (years)	2.249	2.475	12	0
Black player	0.597	0.491		
Recent national team representation	0.223	0.417		
Panel B: Team Characteristics				
Network size	2.211	1.91	6	0
French coach dummy	0.273	0.446		
Mean league appearances in the network‡	16.530	11.551	38	0
League rank	8.370	6.154	20	1
League points	58.507	18.350	95	11
Panel C: Yearly Migration Decisions (%)				
Stay in England	78.0			
Stay in the same English team	63.5			
Observations		422		

†: while not reported, I also control for player positions (e.g., goalkeeper, forward).

‡: this variable is the average league appearances of other French teammates.

In Panel C, I take an initial look at players' yearly migration outcomes. Many players manage to stay in England at the end of the athletic year; while the number of players staying in the same (i.e., current) English team is lower. This indicates that some French players stay in England by transferring to another team in the English Premier League.

4.5.2 Empirical Strategies

I now discuss the empirical strategies for identifying the network effect on yearly migration outcomes. Following Edin et al. (2003) and Munshi (2003) I establish the following linear probability specification:

$$Pr(X_{ij} = 1) = \beta_0 + \beta_1 n_{ij} + \beta_2 stayyr_{ij} + \beta_3' \mathbf{I}_{ij} + \beta_4' \mathbf{T}_{ij} + \delta_j + \varepsilon_{ij} \tag{4.4}$$

where i indexes individual and j indexes year. $X_{ij} = 1$ if i stays in England in year j, and $X_i = 0$ otherwise. X_{ij} can be redefined for other types of outcomes. n_{ij} is the network size, i.e., the number of French teammates in i's team in year j, and $stayyr_{ij}$ is the number of years of previous stay. $\mathbf{I}_{ij}$ is the vector of individual characteristics and $\mathbf{T}_{ij}$ is the vector of team-level characteristics. δ_j is the year dummy and ε_{ij} is the error term.

Economists have long documented the endogeneity problem in studies of network effects on achievement (e.g., earnings, test scores), and one of the most challenging econometric issues is the reflection problem (Manski, 1993). That said, in this paper the reflection problem might only be a minor issue: it exists only if a French player's migration outcome (which determines the plans for his future team) affects his current network. Such possibilities are rare given the fairly short career length in football.

However, there are other possible factors that make ethnic networks endogenous. The major issue is the correlation between omitted variables in ε_{ij} and the network size n_{ij}. The most possible case is that time-varying discrimination (e.g., "domestic bias" in player use) in England or some regions of England might be associated with both French networks and migration outcomes of individual French players. I can partially control for this by including year fixed effects, but omitted variable bias cannot be fully excluded. One solution is to find a variable constructed based on non-England factors that affect immigration of French players to England. I will discuss this solution and the related instrumental variable (IV) identification strategy in the latter section.

4.5.3 The Achievement Variable as the IV, and Its Validity

In this paper, I construct an achievement variable that measures the yearly achievement of the France national team in recent major tournaments upon arrival of each French teammate, and use this constructed achievement variable as the IV for the network size.

For player i with n_i French teammates in year j, each teammate has an individual year of *latest arrival* in England and I denote them as $y_1^{ij}, y_2^{ij}, \cdots, y_{n_i}^{ij}$. Denote the achievement of the France national team as $A(y)$ for the year y. I now construct an achievement variable for individual i (or more precisely, i's French network) in year j by

$$a_{ijn_i} = \frac{A(y_1^{ij}) + A(y_2^{ij}) + \cdots + A(y_{n_i}^{ij})}{n_i} \tag{4.5}$$

As the "average achievement" of the France national team upon arrival of French teammates, a_{ijn_i} serves as the IV for the network size n_i. Note that while

each term on the right hand side is year-specific, the whole achievement variable a_{ijn_i} is not year-specific but player/network-specific.

In practice, $A(\cdot) \in [0, 1]$ can be a normalized variable monotonically decreasing in the rank of the France national team (i.e., increasing in the achievement) in the latest tournament, and lags can also be introduced. As an example, in this paper $A(2001) = 1$ since France won both the 1998 World Cup and 2000 European Championship. In Appendix G of this dissertation I will introduce the construction of $A(\cdot)$ in detail.

Why do we expect this achievement variable to be a valid IV for the network size? In other words, why does the achievement of the France national team upon arrival of each network member predict the network size, and influence the migration outcome only through its impact on the network size? It is easy to understand the first requirement: the achievement of the France national team should be correlated with the network size. The success (or failure) of France is a signal for the quality of French players who play for the national team. However, English teams also purchase French players who are not good enough to play for France, based on their expectation that the skill distribution of all French football player is also correlated with the achievement of France, although later I will argue that this might not always be true.

I now discuss the second requirement, i.e., the exclusion restriction: does the achievement of France affect factors other than the ethnic network? The major threat is that France's achievement might also affect player quality—as anticipated by English teams. I are able to control for player quality by including players' league appearances in the English league to close the channel of player quality through which the achievement of France affects migration out-

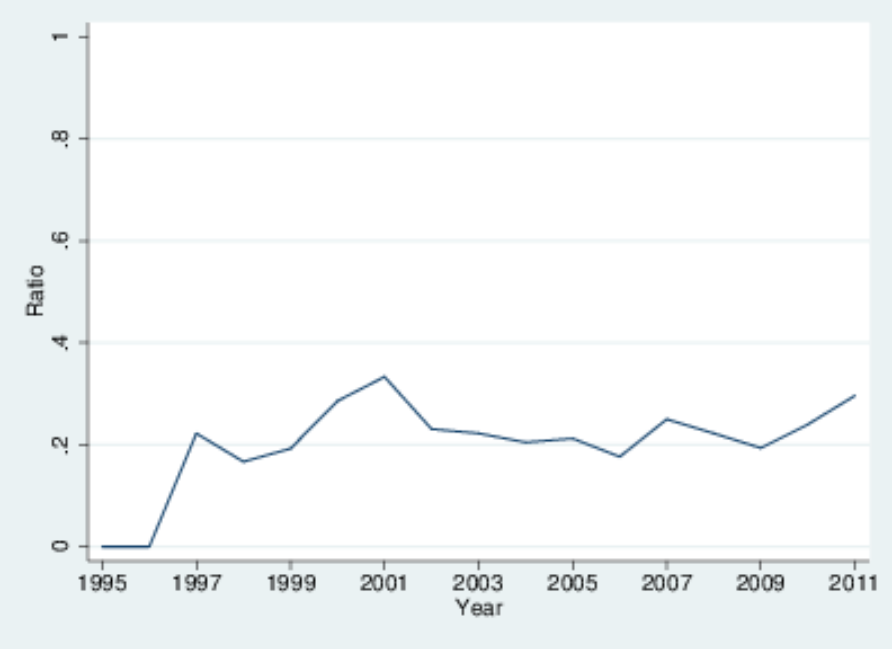


Figure 4.3: The ratio of French players in England who also play for the France national team in the most recent major football tournament for national teams.

comes, but I also argue that the achievement of France is not necessarily related to quality of French players moving in England.

Figure 4.3 shows that in general, only about 20% of all French players in England are able to represent their national team in the most recent national team tournament. The ratio of key players who are major contributors of France's achievement is even lower. Note that the ratio in Figure 4.3 is fairly stable, and in particular, appears to be unaffected by the achievement of the France national team.

While the changing pattern of the skill distribution of top French players can be roughly described based on the achievement of the France national team, it is difficult to determine the skill distribution of all French players migrating to England. Because other major football leagues in Europe (e.g., German Bun-

desliga and Spanish La Liga) also purchase French players, the effect of the achievement of France on the quality of *all* French players in England might actually be minor. In Table 4.2 I use league appearances in England as the proxy for player quality and regress league appearances on the achievement of France upon arrival of players. In Column 1 I regress league appearances only on the achievement of France upon arrival, and find some correlation between France's performance and French players' league appearances. However, such an effect becomes insignificant (and even negative) after including all other control variables, as reported in Column 2. I reconstruct a player-career cross-sectional sample set and and regress the number of average career league appearances in the English league on the achievement of France upon arrival in Column 3, 4, and 5. Again, I find no clear evidence that France's achievement affects individual league appearances in England.

Table 4.2: France's Achievement and League Appearances

	Player-Year		Player-Career		
	(1)	(2)	(3)	(4)	(5)
Achievement	2.994*	−2.277	2.659	3.124	2.607
	(1.529)	(2.157)	(2.428)	(2.420)	(2.470)
Control Variables	No	Yes	No	Yes	Yes
R^2	0.009	0.230	0.009	0.051	0.029
Observations	422	422	141	141	123

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

In column 5 I excludes players who were still playing in England after 2011.

I conduct some further tests in Appendix H of the dissertation. I show that the achievement of the France youth national team, which enrolls more French

players (including many players who will have never played for the adult national team later), appears to be fairly stable in the 1990s and 2000s, and has no effect on French players' league appearances in England. I also regress individuals' average career league appearances on year of arrival fixed effects in player-career cross-sectional data, and still find no significant effect. These results imply that the achievement of the France national team might affect sizes of French ethnic networks in English teams (which can be shown in first-stage regressions), but is probably uncorrelated with player quality, and the IV constructed based on France's achievement should influence migration outcomes only through its impact on the network size.

4.6 Empirical Analysis: Migration Outcomes

In this section, I investigate effects of ethnic social networks on migration outcomes of French football players in the English Premier League. Every summer (which is the end of the current athletic year and the beginning of the new athletic year) football teams and players will decide their future plans, and migration outcomes are thus determined. I study three types of migration outcomes: the outcome of staying in the same English team, staying in the England, and transferring to another team in England.

4.6.1 Reduced-Form and First-Stage Regressions

In Table 4.3, I begin with three reduced-form regressions of migration outcomes on the achievement variable and other regressors. In Column 1 I run the regres-

sion of the outcome of staying in the same team after the current athletic year (i.e., in the new athletic year). The average achievement of France upon arrival of French teammates has no significant effect on whether the player stays in the current team. Average league appearances of French teammates and the presence of a French head coach also have minor effects. Age and previous stay are negatively correlated with the outcome of staying in the same team, and self league appearances are positively related to the probability of stay.

I similarly regress the outcome of staying in England (in either the same team or another English team) and the outcome of moving to another English team in Column 2 and 3, respectively. The achievement variable is similarly unrelated to the probability of transferring to another English team, but is significantly positively correlated with the outcome of staying in England. This indicates that the constructed achievement variable, which is assumed to exogenously affect the network size, is only associated with whether an individual French player stays in England. Its effects on team-level outcomes, however, are minor. Effects of other regressors generally follow their patterns reported in Column 1, but in Column 2 previous stay in England is positively correlated with the outcome of staying in England, and in Column 3 self league appearances are negatively correlated with the outcome of moving to another English team.

In Column 4 and 5 I examine the first-stage relationship between the achievement variable and the network size. In Column 4 I do find that the arrival of a French teammate is largely determined by the achievement of France in recent tournaments. Having a French head coach in the team also makes the French network larger. In general, this first-stage model well predicts the network size and, in particular, the constructed achievement variable is closely related to the

Table 4.3: Reduced-Form and First-Stage Regressions

	Reduced-form: next-year team			First-stage	
Dependent Variables:	Same team	England	Another team	Network Size	
Achievement	0.061	0.168**	0.107	1.385***	1.798***
	(0.088)	(0.078)	(0.072)	(0.217)	(0.159)
Age	−0.188***	−0.112**	0.076	0.012	−0.061
	(0.055)	(0.048)	(0.045)	(0.135)	(0.146)
Previous stay (year)	−0.020*	0.019*	0.038	−0.015	−0.015
	(0.011)	(0.010)	(0.009)	(0.027)	(0.026)
Self league appearances	0.018***	0.015***	−0.003*	0.002	0.003
	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)
Network league appearances	0.001	−0.001	−0.002	0.023***	
	(0.003)	(0.002)	(0.002)	(0.006)	
French coach dummy	0.054	−0.008	−0.061	2.618***	2.639***
	(0.053)	(0.047)	(0.043)	(0.131)	(0.140)
Year fixed effects	Yes	Yes	Yes	Yes	No
Other covariates	Yes	Yes	Yes	Yes	Yes
First-stage F statistic	—	—	—	35.74	61.69
R^2	0.309	0.278	0.161	0.733	0.663
Observations	422	422	422	422	422

Coefficients of selected regressors are reported. The complete table is available upon request.

Dependent variable in (1): staying in the same team in the English Premier League after this year.

Dependent variable in (2): staying in the English Premier League after this year.

Dependent variable in (3): staying in England, but transferring to another English team after this year.

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

network size. In other words, our IV is valid in the sense of the strong first-stage relationship. In Column 5 I drop the average league appearances of teammates and the qualitative pattern remains, although the magnitude of the achievement variable does become larger. I will include this variable of network-level athletic

performance in all following regressions in this paper.

4.6.2 Main Results: OLS and IV Regressions

I now report the main results of this paper: the OLS and IV regressions of three types of migration outcomes on the size of the French ethnic network and other variables. Results are shown in Table 4.4. In Column 1 and 2 I first examine the network effect on whether the player stays in the same English team after the athletic year. Using the OLS model, I find no evidence that the ethnic network is correlated with the outcome of staying in the same English team, although age, previous stay, and self league appearances are all determinants of the probability of staying in the current team, which is consistent with the reduced-form finding in the previous table. I use the network-level achievement variable to instrument for the network size in Column 2, and although I find a positive network effect, such an effect is insignificantly different from zero. Controlling for individual, team, and year characteristics, both the OLS and IV regression show no effect of the ethnic network on the outcome of staying in the same English team for French players playing in the English Premier League. Likewise, while not reported here, no network effects are found in the regression of the dummy of staying in the same English team conditional on staying in England.

In Column 3 and 4 I examine the network effect on the outcome of staying in England using the OLS and IV model. Column 3 does show the positive network effect, and age, previous stay, and self league appearances again influence the migration outcomes, similar to the finding in Table 4.3. However, the OLS estimate of the network effect appears to be downward biased: in Column 4, I

Table 4.4: OLS and IV Regressions: The Effect of Ethnic Networks on Migration Outcomes of French Players

Dependent Variables:	Same team		England		Another English team	
	OLS	IV	OLS	IV	OLS	IV
Network size	−0.004	0.043	0.031*	0.121**	0.034**	0.078
	(0.020)	(0.062)	(0.017)	(0.056)	(0.016)	(0.050)
Age	−0.187***	−0.189**	−0.110**	−0.113**	0.077*	0.075*
	(0.055)	(0.053)	(0.048)	(0.048)	(0.044)	(0.043)
Previous stay (year)	−0.019*	−0.019*	0.020**	0.021**	0.039***	0.040***
	(0.011)	(0.010)	(0.009)	(0.010)	(0.009)	(0.009)
Self league appearances	0.018***	0.018***	0.014***	0.014***	−0.003*	−0.003*
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Network league appearances	0.002	−0.000	0.001	−0.004	−0.001	−0.003
	(0.002)	(0.004)	(0.002)	(0.003)	(0.002)	(0.003)
French head coach dummy	0.061	−0.060	−0.091	−0.325**	−0.152**	−0.264*
	(0.073)	(0.168)	(0.064)	(0.152)	(0.059)	(0.136)
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Other covariates	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.309	—	0.275	—	0.166	—
Observations	422	422	422	422	422	422

Coefficients of selected regressors are reported.

Dependent variable in (1) & (2): staying in the same team in the English Premier League.

Dependent variable in (3) & (4): staying in the English Premier League.

Dependent variable in (5) & (6): staying in England, but transferring to another team.

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

find that in the linear sense, one additional French teammate increases the probability of staying in England after the athletic year by 12.1%; the magnitude of the network effect estimated by the IV model is almost four times greater than the OLS estimate. Note that this also implies that the omitted variable issue (e.g., "domestic bias") is dominant, since the OLS estimate should be upward

biased if the reflection problem is the only source of endogeneity (Manski, 1993). This is consistent with our earlier argument that the reflection problem, which is common in studies of peer effects on achievement, is probably not a serious issue in this paper.

In Column 5 and 6 I analyze if ethnic networks affect the outcome of transferring to another English team. If, as predicted by many economists (e.g., Hoxby, 2000), ethnic networks help immigrants assimilate, then I might be able to observe the positive network effect on the outcome of staying in England but moving to another English team given that ethnic networks fail to help French players stay in the same team. In Column 5 I do find the significantly positive OLS estimate of the network effect. However, consistent with the non-significant reduced-form result reported previously, the IV regression in this table implies that the ethnic network does not help a French player moves to another English team. Of course, most prior findings of the network effect on immigrants' assimilation mainly focus on low-skilled (Damm, 2009) and adolescent (e.g., Hoxby, 2000; Stiefel et al., 2004) immigrants, and the conclusion does not necessarily apply to immigration of high-skilled professionals.

4.6.3 Additional Tests: Heterogeneous Effects

I finally conduct three sets of additional tests on heterogeneous network effects and conclude the empirical part of this paper. In Table 4.5 I split French players in the sample into various subpopulations by demographic characteristics, including age and previous stay in England. Generally, age at arrival is closely related to the degree of assimilation (e.g., Stiefel et al., 2004; Bleakley and Chin,

2010), and younger French players should adapt to the football environment in England faster.

Table 4.5: Heterogeneous Network Effects: by Demographic Characteristics

Dependent Variables:	Same team		England		Another Eng. team		Sample
	OLS	IV	OLS	IV	OLS	IV	Obs.
Age < 25	0.015	0.144	0.015	0.440	−0.000	0.108	149
	(0.037)	(0.131)	(0.032)	(0.620)	(0.030)	(0.106)	
Age < 30	−0.002	0.103	0.037*	0.198***	0.039*	0.095*	316
	(0.023)	(0.067)	(0.020)	(0.062)	(0.019)	(0.054)	
Age ≥ 25	0.009	0.051	0.046**	0.108	0.037*	0.057	273
	(0.025)	(0.077)	(0.022)	(0.067)	(0.021)	(0.063)	
Age > 28	0.022	−0.008	0.041	−0.047	0.019	−0.039	145
	(0.036)	(0.113)	(0.029)	(0.095)	(0.030)	(0.096)	
Previous stay ≤ 2	0.014	0.077	0.015	0.102*	0.001	0.025	274
	(0.023)	(0.066)	(0.021)	(0.061)	(0.016)	(0.047)	
Previous stay > 2	−0.032	−0.081	0.084**	0.208*	0.116***	0.289**	148
	(0.043)	(0.136)	(0.034)	(0.112)	(0.037)	(0.128)	
Previous stay > 3	−0.041	−0.365	0.082*	0.100	0.123**	0.465*	104
	(0.055)	(0.278)	(0.041)	(0.174)	(0.049)	(0.266)	

All regressors and year fixed effects are included. Coefficients of *network size* are reported.

All first-stage regressions have R^2 no less than 0.7 and F-statistic greater than 10.

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

As expected, in Row 2 I do find that ethnic social networks have significantly positive effects on outcomes of staying in England as well as transferring to another English team among players under 30. The network effect on the outcome of staying in England is particularly large. The assimilation effect is a possible mechanism through which the ethnic social network helps immigrant players since players are able to move to another team in England even if failing to stay

in the current team. However, results of this table show no network effect on other age groups, and in particular, those who are older than 28 (and even 25). In other words, veteran players—who should probably feel more difficult to assimilate—receive no significant benefit of any kind from French ethnic social networks.

In the last three regressions I focus on three subpopulations split by previous stay in England. Having a larger network does increases the overall probability of staying in England for French players who have played in England no more than two years. On the other hand, however, the network effect on those who have stayed in England more than two years is almost doubled. Although ethnic networks does not help those with more experiences in England stay in the same team, the likelihood of moving to another English team and the overall likelihood of staying in England are positively correlated with the network size. In general, network effects appear to be relatively weaker among French players who are not well experienced in England.

In Table 4.6 I turn to focus on subpopulations split by athletic performance. Ethnic social networks are found to help workers with lower skills in the labor market (e.g., Edin et al., 2003; Munshi, 2003). Athletic performance is closely related to football skills and are measured in two ways in this table. The first, *pre-measured* way relies on whether the player represented the France national team in the latest tournament for national teams. The second, *post-measured* way is based on league appearances in the current year. Basically, players with latest France representation and with more league appearances are expected to have "better" skills. In the first two regressions I find that ethnic networks have significantly positive effects on French players who did not represent France.

These players are also more likely to stay in England by moving to another English team, with the help from ethnic networks. However, there is no significant network effect on the outcome of staying in the same team after the athletic year. Ethnic social networks appear to be unrelated to French players with latest France representation.

Table 4.6: Heterogeneous Network Effects: by Athletic Performance

Dependent Variables:	Same team		England		Another Eng. team		Sample
	OLS	IV	OLS	IV	OLS	IV	Obs.
Latest France rep. = 0	−0.013	0.063	0.019	0.158**	0.031*	0.095*	328
	(0.022)	(0.068)	(0.020)	(0.064)	(0.019)	(0.057)	
Latest France rep. = 1	−0.018	0.063	0.054	−0.143	0.072*	−0.206	94
	(0.050)	(0.159)	(0.041)	(0.148)	(0.036)	(0.156)	
League appearance ≤ 13	−0.037	−0.003	−0.015	0.157	0.021	0.160	131
	(0.040)	(0.164)	(0.040)	(0.177)	(0.031)	(0.138)	
League appearance ≤ 26	−0.016	0.090	0.003	0.164*	0.019	0.074	258
	(0.028)	(0.099)	(0.024)	(0.092)	(0.023)	(0.080)	

All regressors and year fixed effects are included. Coefficients of *network size* are reported.

All first-stage regressions have R^2 no less than 0.7 and F-statistic greater than 10.

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

In the last two regressions I study players with no more than 13 (1/3 of all matches in a year) and 26 (2/3 of all matches) league appearances. Ethnic networks appear to offer no help for French players with few league appearances to extend their stay in England. I do find that ethnic networks increases the probability of staying in England among players who have appeared no more than two thirds of all league matches; however, there is still no network effect on the outcome of staying in the same team. Results in Table 4.6 show that ethnic networks do not really positively affect migration outcomes of French player-

s with lower levels of athletic performance—and these players are those who need the support from ethnic social networks most.

In Table 4.7 I split the sample by team characteristics. In the first three rows I split the population by team rankings. Although I again find no network effect on the outcome of staying in the same team, ethnic networks do help extend the stay in England for French players in all tiers of English teams. The network effects among players in top four teams—which qualify for the European Champions League—are very large, but ethnic networks also significantly help French players in lower-tier teams in England.

The fourth regression studies players in teams where head coaches are not French. Intuitively, the presence of the French coach reduces domestic bias and helps French players adapt to the environment in England, while an ethnic network can be viewed as the compensation of the absence of the French coach. I do find some positive network effects on the outcome of staying in England, although the main contribution of ethnic networks is to help migrant players stay by moving to another team.

The last three regressions study network effects in teams with various network sizes. I first focus on French players with no and only one French teammate. Results show that compared those with no French teammate, having one compatriot teammate greatly increases the probability of staying in England; while the one-teammate network does not affect the likelihood of staying in the current team, it does help a French player stay in England by moving to another English team. I then extend the analysis and focus on French players playing for teams where there are no more than two compatriot teammates (note that the average network size is slightly greater than 2). Albeit still significantly pos-

Table 4.7: Heterogeneous Network Effects: by Team Characteristics

Dependent Variables:	Same team		England		Another Eng. team		Sample
	OLS	IV	OLS	IV	OLS	IV	Obs.
Rank ≤ 4	−0.050	0.306	0.008	0.343*	0.058*	0.037	172
	(0.046)	(0.204)	(0.036)	(0.180)	(0.034)	(0.133)	
Rank > 4	0.009	0.027	0.059**	0.134**	0.050**	0.107**	250
	(0.026)	(0.065)	(0.024)	(0.061)	(0.021)	(0.053)	
Rank > 10	−0.014	0.083	0.020	0.158*	0.035	0.076	157
	(0.045)	(0.093)	(0.041)	(0.086)	(0.036)	(0.073)	
No French head coach	−0.009	0.020	0.050**	0.102*	0.059***	0.082*	307
	(0.025)	(0.060)	(0.022)	(0.053)	(0.021)	(0.049)	
Network size ≤ 1	−0.032	0.022	0.150	0.229*	0.182**	0.207**	190
	(0.109)	(0.128)	(0.101)	(0.120)	(0.084)	(0.099)	
Network size ≤ 2	0.023	0.063	0.081*	0.146*	0.059	0.084	271
	(0.052)	(0.094)	(0.046)	(0.082)	(0.041)	(0.073)	
Network size ≥ 2	−0.012	0.280	−0.002	0.121	0.010	−0.158	232
	(0.030)	(0.197)	(0.025)	(0.142)	(0.026)	(0.154)	

All regressors and year fixed effects are included. Coefficients of *network size* are reported.

All first-stage regressions have R^2 no less than 0.5 and F-statistic greater than 10.

Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

itive, the network effect on the outcome of staying in England becomes smaller, and I now find no effect on the outcome of moving to another English team. While not reported here, I observe that the magnitude of the network effect is decreasing in network size. The final regression studies the network effect among French players playing with at least two French teammates, and I find no network effect of any kind. These regressions imply that social networks have greater effects among French players in smaller ethnic networks, and the marginal network effect is decreasing.

4.7 Conclusion

This paper examine social network effects on yearly migration outcomes of French football players playing in the English Premier League. I investigate migration outcomes after an athletic year (or a *season*) along three dimensions: the outcome of staying in the same (current) English team, the outcome of staying in England, and the outcome of staying in England by moving to another team in the English league.

The major methodological challenge of the empirical analysis of the effect of ethnic social networks is the problem of endogenous networks. The traditional reflection problem (Manski, 1993) implies that individual achievement can be affected by, but also reversely affect peer achievement. While in this paper, I see the reflection problem as a minor threat (since an individual's migration outcomes related to his future plans is unlikely to affect the ethnic makeup in the current team), ethnic networks might still be endogenous since there are omitted factors affecting both networks and migration outcomes, such as domestic bias of head coaches and teams. I construct a variable that reflects the achievement of the France national team upon arrival of each French teammate (i.e., each network member) and use this achievement variable to instrument for the size of the ethnic social network. I find that France's achievement predicts the influx of French players, and use various information and approaches to ensure that France's achievement influences migration outcomes exclusively through its impact on the network size.

I find the general network effect on the overall outcome of staying in England. In many subpopulations I also observe that a larger ethnic network is

associated with a higher probability of staying in England by transferring to another English team. However, in all regressions I find no network effect on the outcome of staying in the current English team. I also observe heterogeneous network effects among French players in the English Premier League. I find particularly large network effects among several specific subpopulations, including younger (under 30) players, players who have stayed in England for more than two years, players who are not national team members, and players in small ethnic networks. However, ethnic networks do not always help those in most need extend their stay in England: French players who are older or have few league appearances appear to receive no significant benefit from ethnic networks.

This paper provides new evidence for the literature of social networks and high-skilled immigration. Previous studies on sport labor economics mainly focus on market effects, while this paper focuses on impacts of the social network, which is considered to be "non-market". From the perspective of economics of immigration, this is surely not the first paper about network effects on immigrants; however, many studies investigate low-skilled immigrants from developing economies and show robust results of significant network effects. In this paper, however, our empirical findings suggest that the traditional conclusion on social networks and immigration should be extrapolated to high-skilled professionals with caution: indeed, I observe significant network effects only on some types of migration outcomes among some subpopulations. This is somewhat similar to the conclusion of Freeman and Huang (2015) that the network effect among high-skilled professional immigrants are not always positive and it depends on certain contexts.

This paper follows the standard econometric setting in labor economics (e.g., Edin et al., 2003; Munshi, 2003) and measures the strength of a network using its size. I also only consider networks to be unweighted and non-directed. A more complicated network structure might lead to new findings for both theoretical and empirical research. Another avenue of future exploration involves the mechanisms behind network effects. Immigrant can acquire job skills from network members (the learning effect), or just learn how to live in the host country more comfortably (the assimilation effect). Can we decompose the network effect? This requires new data sets and further empirical discussions.

CHAPTER 5
CONCLUDING REMARKS

This dissertation studies social networks from three main aspects: the formation of the social network (e.g., Marmaros and Sacerdote, 2006), the representative characteristics of the social network (e.g., McPherson et al., 2001), and the social and economic effects of the social network (e.g., Granovetter, 1973; Montgomery, 1991). In this dissertation, I provide thorough discussions of the above three aspects and present three case studies in the context of international migration.

Compared with the general population, immigrants similarly rely on social networks—if not more. By sharing information and provide support within the network, immigrants' social and economic outcomes can be significantly improved. On the other hand, staying in social networks also has some clear disadvantages and might hurt immigrants. This leads to the natural questions regarding immigrants' social networks, which are discussed in this dissertation: Why do some immigrants form social networks? If immigrants' social networks are formed, what kinds of representative characteristics of social networks can be observed? Finally, what are the social and economic consequences of being in social networks for immigrants? In Chapter 2, 3, and 4, I present three empirical case studies that are related to each of the three topics, respectively.

Immigrants' social networks can be formed in a very trivial way: immigrants like to be bond with others. However, social scientists are usually interested in more complicated mechanisms that possibly drive the social phenomena. In Chapter 2, I use American Community Survey (ACS) data and study the formation of a particular type of the need-based social network—carpooling networks

for immigrants who commute to work. In this context, the main reason behind network formation is not (or at least not only) that immigrants want networks, but immigrants *need* networks: I argue that English proficiency is a key factor affecting immigrants' tendency to carpool with others. Specifically, immigrants who have higher proficiency in English should be less likely to experience the difficulties related to language when driving (such as confusion of navigation, issues with traffic police, or language-related psychological concerns), and thus are less likely to carpool. In the empirical analysis, I use the interaction term between age at arrival and country of origin to instrument for English proficiency (Bleakley and Chin, 2004, 2010; Guven and Islam, 2015) and thus separate out the cultural and linguistic effect. I find the causal relationship between language proficiency and carpooling behaviors: immigrants who speak English better are indeed less likely to carpool, and furthermore, have fewer co-riders when commuting to work.

Researchers have long observed that many social networks consist of co-ethnic members and can thus be defined by the demographic characteristics. One concern is that immigrants do not interact with every person in the ethnic group, and if "sub-networks" are formed, what are the possible representative characteristics of such sub-networks? In Chapter 3, I use online social networking data retrieved from Renren.com and study homophily based on English-name usage of Chinese graduate students in the U.S. I argue that English-name usage is a typical behavior of acculturation (Gordon, 1964), and thus some ethnic social networks—in which members want to acculturate—stem from acculturational homophily. In Chapter 2, I exploit a natural linguistic experiment on English-name usage: Chinese students whose original Chinese names are difficult to be pronounced by native speakers of English are more likely to be

English-name users. Employing this approach, I find that individual English-name usage is causally related to English-name usage of close friends. This indicates that after excluding the possibility that the self behavior is influenced by others, the individual acculturational effort still leads to the representative characteristic of the whole social network.

The final case study concerns how ethnic social networks affect migration decisions of highly professional immigrants. In Chapter 5, I focus on professionals who travel between two developed countries and work in a highly globalized labor market, namely, French football (*soccer*) players in the English Premier League. Defining the ethnic social network by compatriot (French) teammates, I find that the French network generally does not have the significant effect on the outcome of staying in the current team, but does improve the likelihood of staying in England for French players. The network effects, however, are highly heterogeneous, in the sense that not every player is affected by the network. Specifically, ethnic networks appear not to always help those "most in need" for support for stay, such as veteran players or players with relatively low outputs. The results indicate that traditional findings of ethnic social networks among low-skilled immigrants (e.g., Munshi, 2003; Damm, 2009) must be extrapolated with caution when discussing other types of immigrant populations (e.g., Garip, 2012), such as high-skilled immigrants.

As stated earlier, this dissertation contributes to the existing literature by studing all three main topics of social network research in the context of immigration, and in particular, with the foci on specific immigrant sub-populations. I finally discuss how it leads to potential research in future briefly. The main data issue of this dissertation is that it cannot rely on longitudinal data to study

social networks in the time dimension. Having the panel structure might allow us to decompose the individual-level and network-level effects and help us understand the mechanisms behind social phenomena related to networks. For example, with longitudinal data it is likely to separate out the contribution of peer selection and peer influence on the presence of behavioral homophily, if the time variation account for one channel exactly. The disseration also points out another possible avenue for future work, namely, to analyze the structure of the social network. Specifically, it demonstrates that with the variation of strength among network members, their influences also vastly vary. The above findings of this dissertation can thus lead to future theoretical frameworks and empirical investigations.

29 August, 2016

Zagreb, Republic of Croatia

APPENDIX A

THE OVERVIEW OF THE APPENDIX

This part provides an overview of the structure of the section Appendix. The first case study does not have any appendix. From Appendix B to E, I present four additional sections for the third chapter, *Acculturational Homophily*. In Appendix F I present another case study that is the extended paper based on the idea and methodology of Chapter 3. This case study focuses on efforts for cultural assimilation and graduate school choices, which are related to the materials presented in the main part of the dissertation. Note that although there are some overlapping materials in Chapter 3 and Appendix F, I still present the full content of the paper in Appendix F to keep its completeness. Appendix G and H are two additional sections for the third chapter, *Social Networks and Immigration of High-Skilled Professionals*.

APPENDIX B

APPENDIX FOR CHAPTER 3, PART A: SCHOOL TIERS

In this appendix I discuss how the tiers of colleges and graduate schools are classified. I split both Chinese colleges and U.S. graduate schools into three tiers based on school rankings and reputation-based school alliances.

For Chinese colleges, tier 1 colleges include members of the "C9 League". Equivalent to Germany's *Exzellenzinitiative*, Chinas C9 League comprises nine most renowned universities in Mainland China[1]. Tier 2 colleges include all other universities sponsored by "Project 985"[2], which is an official project initiated by national and local governments that allocate funding to 39 reputable research universities in Mainland China. Tier 3 colleges include all other Chinese schools.

For U.S. graduate schools, tier 1 schools include universities in top 10 of the US News Best Global University Rankings, plus all other Ivy League schools[3]. Tier 2 schools include all other members of the Association of American Universities (AAU), which is an organization of 62 leading research universities in North America. Tier 3 schools include all other U.S. universities in the sample.

[1] This C9 League contains Peking University, Tsinghua University, University of Science and Technology of China, Fudan University, Nanjing University, Shanghai Jiao Tong University, Zhejiang University, Harbin Institute of Technology, and Xian Jiao Tong University.

[2] All C9 League members (tier 1 colleges) are sponsored by this project.

[3] Based on these criteria, tier 1 schools include Harvard University, Massachusetts Institute of Technology, University of California-Berkeley, Stanford University, California Institute of Technology, University of California-Los Angeles, University of Chicago, Yale University, Columbia University, University of Pennsylvania, Cornell University, Brown University, and Dartmouth College.

APPENDIX C

APPENDIX FOR CHAPTER 3, PART B: THE IDENTIFICATION OF DIFFICULT-TO-PRONOUNCE CHINESE NAMES

In this paper, I exploit a natural experiment on English-name usage which relies on the classification of Chinese names by the pronunciation difficulty in English. In this appendix, I introduce the criteria of identifying "difficult-to-pronounce" names. The pronunciation difficulty exists generally due to the huge linguistic difference between two languages (Crowley and Bowern, 2010). Although the *pinyin* system of Chinese romanizes Chinese characters into the Latin alphabet, the system cannot always reflect the pronunciation rules of Chinese precisely (Bassetti, 2007).

Table C.1: Typical Difficult-to-Pronounce Phonological "Blocks" in Chinese

Phonological "block" in Chinese	Syllable example	Actual approx. syllable in English	Ideal approx. syllable in English	Wade-Giles character
−ang	*sh**ang***	*shan*	*shawn*	*shang*
ca−	***ca**n*	*kan*	*tsan*	*ts′an*
ce−	***ce**n*	*sen*	*tsen*	*ts′en*
co−	***co**ng*	*kong*	*tsong*	*ts′ung*
cu−	***cu**n*	*kun*	*tsun*	*ts′un*
−eng	*sheng*	*shen*	*shewn*	*sh3ng*
−he	***he***	*hi*	*her**	*ho*
−i†	*ch**i***	*chi*	*ch*‡	*ch′ih*
−ue	***yue***	*you*	*jüe*◇	*yüeh*
x−	***x**u*	*zu*	*schü*	*hsü*
yu	***yu***	*you*	*ü*	*yü*

†: Only if *−i* follows *c*, *r*, *s*, *z*, *ch*, *sh*, *zh*. ∗: In British English (i.e., non-rhotic for *−er*). ◇: In German.

‡: The sound is like the consonant but the syllable is pronounced as if it is a vowel.

There are three categories of phonological "blocks" in Chinese that are difficult to pronounce in English. In this paper, I consider eleven "difficult-to-pronounce" phonological blocks in total. In Column 1 and 2 of Table C.1, I list these blocks and examples of the syllables. These syllables are pronounced differently in English, and I show the actual approximate syllables of English pronunciation in Column 3. To show the difference, in Column 4 I show how these syllables should be pronounced in English ideally. I finally present the corresponding Wade-Giles characters as the reference in Column 5. Similar to *pinyin*, Wade-Giles is a romanization system for Chinese but has been replaced in Mainland China for decades. As the system is invented by native speakers of English, it reflects how to use English to pronounce Chinese syllables better than *pinyin*, but still not completely precisely alike.

Among three categories of "difficult-to-pronounce" blocks, the first category involves the difference between phonological blocks of the velar nasal and the alveolar nasal (e.g., Zee, 1985; Lee and Zee, 2003), including *-ang* and *-eng*. I do not include *-ong* and *-ung* as there are no *-on* and *-ung* in Chinese. I also exclude *-ing* as the difference between *-in* and *-ing* is arguably much smaller.

The second category involves the phonological block *x-*, which is widely used in Chinese but relatively uncommon in English. Hence native speakers of English usually find it difficult to pronounce *x-*, and the pronunciation of *x-* in Chinese is substantially different from its common approximate syllable in English.

The third category involves phonological blocks that are widely used in both languages, but have different pronunciation rules. All blocks with *c-* in English are included in the list. For example, *co-* is pronounced as *ko-* in English and *tso-*

in Chinese. Other similar blocks, including *-he*, *-i* (in some cases), *-ue*, and *yu*, are also presented in Table C.1.

APPENDIX D

APPENDIX FOR CHAPTER 3, PART C: THE PRONUNCIATION DIFFICULTY AMONG NON-MIGRANTS

It will be a concern if the pronunciation difficulty is a special feature among non-migrant students who have no plan to receive education abroad. In other words, the identification strategy should be more convincing if non-migrant students stay in China not because that their Chinese names are difficult to pronounce.

Due to data limitation, I am unable to examine a similar representative samples of non-migrants retrieved from Renren. Most schools in China do not provide publicly available student lists, and it is even more difficult to find alumni records that contain both names and post-graduation outcomes. In this paper I find two samples of graduates from China Center for Economic Research (CCER) at Peking University and Nanjing Foreign Language School (NFLS) that provide both students' names and their post-graduation information (whether staying in China or moving abroad). I retrieve 18 sub-samples in total and the two-digit number following the name of the institution reflects the class. In Table D.1, I summarize the distribution of difficult-to-pronounce names among both migrant and non-migrant students in these student samples.

Results show that: (a) in general, migrant students appear to be slightly more associated with the pronunciation difficulty, but the difference between migrant and non-migrant students is statistically insignificant; (b) the average proportion of difficult-to-pronounce names in these external samples is very close to that in the sample used in this paper, namely around 42% and 43%. This implies that the pronunciation difficulty is probably not a unique characteristic for a specific population, and there is no clear evidence of selective migration based

Table D.1: The Pronunciation Difficulty in External Data with Non-Migrant Students

	Domestic (non-migrant)			Abroad (migrant)		
Sample	Total # of students in the sample	# of students with diff.-to-pro. names	Ratio	Total # of students in the sample	# of students with diff.-to-pro. names	Ratio
CCER 04	271	114	41.9%	56	19	33.9%
CCER 05	406	156	38.4%	82	38	46.3%
CCER 06	406	171	42.1%	82	36	43.2%
CCER 07	404	167	41.3%	109	46	42.2%
CCER 08	299	132	42.1%	90	37	41.1%
CCER 09	447	186	41.6%	108	48	44.4%
CCER 10	518	224	43.2%	144	62	42.4%
CCER 11	590	247	41.7%	182	76	41.5%
CCER 12	392	167	42.7%	181	76	41.8%
CCER 13	407	175	42.8%	164	72	43.4%
NFLS 07	275	119	43.2%	174	74	42.0%
NFLS 08	152	55	43.4%	214	98	45.8%
NFLS 09	220	92	41.8%	231	102	44.1%
NFLS 10	187	80	42.8%	242	100	41.3%
NFLS 11	208	92	44.2%	270	115	42.6%
NFLS 12	208	88	42.3%	262	111	42.4%
NFLS 13	194	81	41.8%	264	122	46.2%
NFLS 14	210	101	48.1%	261	117	44.8%
Pooled	5,794	2,458	42.4%	3,116	1,349	43.3%

Observations: 5,338 (CCER samples); 3,572 (NFLS samples); 8,910 (all samples).

on pronunciation issues in China.

APPENDIX E

APPENDIX FOR CHAPTER 3, PART D: ADDITIONAL TESTS OF ACCULTURATIONAL HOMOPHILY

In this appendix, I conduct several additional tests to check the robustness of the main results. In Table E.1 I use two alternative measures of the extent of acculturational homophily: the percentage of close friends with English-name usage, and whether there is at least one close friend who uses the English name. These measures are defined conditional on listing close friends on Renren, and in Table E.1 I only focus on the sub-sample that excludes students who present zero close friend.

Table E.1: Additional Tests: Other Measures

	% with English-name usage		If > 0 with English-name usage	
	(1)	(2)	(3)	(4)
	OLS	IV	OLS	IV
English-name usage	0.150***	0.224***	0.414***	0.590***
	(0.008)	(0.041)	(0.018)	(0.096)
# of close friends	−0.002	−0.002	0.040***	0.038***
	(0.001)	(0.001)	(0.003)	(0.003)
Individual characteristics	Yes	Yes	Yes	Yes
School covariates	Yes	Yes	Yes	Yes
Pre-arrival characteristics	Yes	Yes	Yes	Yes
Post-arrival characteristics	Yes	Yes	Yes	Yes
Sample	Sub	Sub	Sub	Sub
Observations	3,171	3,171	3,171	3,171

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In Column 1 of Table E.1 I run an OLS regression of the percentage of close friends with English-name usage on self English-name usage and other control variables. I observe the presence of acculturational homophily using this new measure. Furthermore, using the pronunciation difficulty to instrument for English-name usage in Column 2, I find that using the English name increases the percentage of close friends who are English-name users by more than 20%.

In Column 3 and 4 I repeat the exercise in Column 1 and 2, but now using the indicator of having at least one close friend who uses the English name as the dependent variable. Both the OLS and IV model indicate that self English-name usage appears to be positively correlated with the likelihood of having close friends who are English-name users, and in Column 4 the IV model estimates that having the English name increases the likelihood of having at least one close friend with English-name usage by almost 60%.

In Table E.2 I reexamine the main results presented in the main part of the paper (Table 3.9). However instead of using school tier fixed effects, I run four regressions with college and graduate school fixed effects that control for all unobservable university-specific characteristics.

In the first two columns I estimate the effect of self English-name usage using OLS and 2SLS in the full sample. I again show the presence of acculturational homophily, and its extent is quantitatively close to that reported in Table 3.9. The similar pattern is observed in Column 3 and 4, where I investigate the subsample that excludes students who do not show close friends online. In general, additional tests in this appendix indicate the robustness of the main results of this paper, in the sense that acculturational homophily can be similarly observed using other measures, and school fixed effects models.

Table E.2: Additional Tests: School Fixed Effects

	# of close friends with English-name usage			
	(1)	(2)	(3)	(4)
	OLS	IV	OLS	IV
English-name usage	0.289***	0.443***	0.576***	0.943***
	(0.012)	(0.065)	(0.025)	(0.137)
# of close friends	0.062***	0.061***	0.053***	0.050***
	(0.002)	(0.002)	(0.004)	(0.004)
Control variables	Yes	Yes	Yes	Yes
Sample	Full	Full	Sub	Sub
First-stage F-statistic	—	213.37	—	87.35
Observations	7,222	7,222	3,171	3,171

Standard errors are in parentheses. *: $p < .05$; **: $p < .01$; ***: $p < .001$.

APPENDIX F

AN EXTENDED STUDY BASED ON CHAPTER 3: EFFORTS FOR CULTURAL ASSIMILATION AND GRADUATE SCHOOL CHOICES: ACADEMIC PURSUITS VERSUS LOCATIONAL PREFERENCES?

F.1 Abstract of the Study

Using social networking data, this paper studies the effects of efforts for cultural assimilation on Chinese students' school choices when applying for U.S. graduate schools. I use English-name usage to measure assimilation efforts among Chinese students. The identification strategy is based on a natural experiment: the difficulty of pronouncing the original Chinese name in English exogenously determines English-name usage. I find that, overall, there is no effect of English-name usage on the tier of the U.S. graduate school attended. However, English-name usage affects the interaction between the school tier and local demographic characteristics: English-name usage is positively associated with attendance of top-tier schools in areas that are traditionally "less chosen" by Asian immigrants, which are defined based on the local racial makeup. The results suggest the possible role of cultural assimilation in making joint school-location choices when students take both academic pursuits and locational preferences into consideration.

F.2 Introduction

Since Chiswick's early work (1978, 1980), economists have long discussed the socioeconomic consequences of Americanization. Cultural assimilation (Gordon, 1964) is a crucial component of Americanization and is recognized as a major factor affecting immigrants' lives (e.g., Tainer, 1988; Borjas, 1994; Bleakley and Chin, 2004, 2010). Moreover, recent studies show that even pre-migration characteristics can influence immigrants' lives after arrival (e.g., Feliciano, 2005; Levels et al., 2008), as such characteristics are highly related to post-arrival assimilation.

The assimilation theory (e.g., Gordon, 1964; Portes and Zhou, 1993) considers language attainment as a good measure of cultural assimilation[1]. It is, however, difficult to study the effect of language attainment on academic outcomes among highly educated immigrants because most of them have acquired professional proficiency in English even long before moving to the U.S.

Alternatively, researchers consider local-name usage as a typical effort for cultural assimilation (Larkey et al., 1993; Shifman and Katz, 2005; Rubinstein and Brenner, 2014). There does exist some heterogeneity in English-name usage among immigrants (e.g., Sue and Telles, 2007; Edwards and Caballero, 2008; Abramitzky et al., 2014) even after controlling for educational attainment. From this perspective, I study the effect of efforts for cultural assimilation on academic outcomes. Specially, using online social networking data retrieved from Renren.com, I examine the relationship between English-name usage and graduate school admission results among students who receive undergraduate education

[1]For example, language skills can be acquired through pre-migration human capital investments, which affect immigrants' post-migration educational outcomes (e.g., Chiswick and Miller, 1994).

in China and graduate education in the U.S. I focus on rankings of graduate schools, which are categorized into school *tiers* in a coarse manner. This paper adds to the literature by empirically linking naming assimilation with school choices.

Understanding the relationship between naming assimilation (and more broadly, efforts for cultural assimilation) and immigrants' academic outcomes also has policy implications. The U.S. is the major destination for international high-skilled workers (Freeman, 2006), and its economy also benefits from ethnic diversity (e.g., Ottaviano and Peri, 2005, 2006; Peri, 2012). It is thus useful to understand determinants—including assimilation efforts—of immigrants' various social outcomes. Individual school choices might also have various types of regional impacts (e.g., Urquiola, 2015).

It is methodologically difficult to identify the causal effect of English-name usage due to the endogeneity problem. The tier of the school attended might reversely affect English-name usage. Moreover, actual English-name usage might be mis-measured based on online data. To tackle these problems, I construct an instrumental variable (IV) approach based on a language-related natural experiment: due to differences in phonological properties between Chinese and English (e.g., Mok and Dellwo, 2008), there are syllables in Chinese (and furthermore, names containing such syllables) that are difficult to pronounce in English. I show that the "pronunciation difficulty" is almost randomly "assigned" in the sample and is unrelated to individual characteristics, and students with difficult-to-pronounce names are indeed more likely to use English names.

I employ this method to examine the effect of English-name usage on school choices. In short, the effect is heterogeneous among Chinese graduate students.

There is no significant effect of English-name usage on the tier of the school attended overall, but English-name usage is positively correlated with the school tier among female students. Moreover, English-name usage affects the interaction between school tiers and local demographic characteristics. English-name usage is positively associated with top-tier graduate school attendance in areas with low Asian or White populations, or high Black populations—such areas are considered to be traditionally "less chosen" by Asian immigrants. Again, although overall English-name usage affects students' joint school-location outcomes, its effect is much greater among female students. The results point out the possible role of English-name usage in making joint school-location choices among Chinese students, and suggest that efforts for cultural assimilation might affect graduate school admission results when locational preferences are taken into consideration.

The remainder of the paper is structured as follows. Section F.3 introduces the background of the paper. Section F.4 introduces data and methods. Section F.5 reports the findings and discusses the results. Section F.6 concludes the paper.

F.3 Background

This section introduces the background of this paper. I first discuss local-name usage as a typical form of cultural assimilation. I then discuss the determinants of English-name usage among Chinese students. Subsequently, I introduce how school choices are evaluated in this paper. I finally discuss the effect of English-name usage on school choices among Chinese students.

F.3.1 English-Name Usage

Social scientists have long studied why and how immigrants culturally assimilate into the mainstream society, and what are its consequences. Immigrants might culturally assimilate into the host society simply for the purpose of convenience. However, as many assimilation efforts—such as language attainment—require human capital investment and are thus costly, there should be a benefit-cost mechanism behind assimilation. For the specific case of linguistic assimilation, the benefits of language attainment are clear: English proficiency is correlated with earnings (e.g., McManus et al., 1983; Tainer, 1988; Bleakley and Chin, 2004), as well as many other outcomes, such as residential choices and intermarriage (Meng and Gregory, 2005; Bleakley and Chin, 2010).

There are some other types of assimilation efforts that are also possible determinants of immigrants' socioeconomic status but, at first glance, do not require human capital investment and appear to be less costly. Immigrants might give up attachment to own cultural traditions that are negatively correlated with labor market outcomes (e.g., Bisin et al., 2011). Specifically, local-name usage affects earnings (e.g., Arai and Thoursie, 2009) but does not require any investment. The reason why some immigrants do not make these "costless" efforts is that such efforts are *not* costless. In this case, local-name usage is related to ethnic and cultural identities (Nicoll et al., 1986; Larkey et al., 1993), and giving up original identities leads to the loss of social capital from ethnic networks (e.g., Portes and Zhou, 1993), which can hurt immigrants (e.g., Munshi, 2003; Damm, 2009, 2014). Hence there is a similar benefit-cost mechanism behind local-name usage. Indeed, immigrants and minorities receive benefits from local-name usage because it effectively avoids name-based discrimination (e.g., Bertrand and

Mullainathan, 2004; Ahmed and Hammerstedt, 2008; Oreopoulous, 2011; Zussman, 2013), as discrimination against minorities is widely seen (e.g., Carlsson and Rooth, 2007; Drydakis, 2012). Such benefits compensate for potential losses of local-name usage.

Similar to other migrants, some Chinese graduate students are also English-name users. In fact, English-name usage is even a "tradition" of English education in China: Chinese students generally start learning English in primary school and English is a major subject throughout secondary and post-secondary education in China[2]. English-name adoption is an effective approach for language teaching (Edwards, 2006), and is especially popular in China (Gao et al., 2005).

This suggests that Chinese students are likely to start using English names even long before receiving education abroad. Since the process of preparation for standardized tests (such as TOEFL and GRE) is long, students usually need to plan graduate studies early[3], and English-name usage reflects an early effort for future cultural assimilation. Education researchers find that English-name usage is related to changes in attitudes towards cultural identities among college students in China (Gao et al., 2005), and students actively experience self-identity changes to prepare for assimilation after arrival. From this perspective, English-name usage serves as the proxy for efforts for cultural assimilation.

Although most Chinese students have adopted English names through language learning, there might still exist variation in English-name *usage* since it is

[2]English is a mandatory subject in China's National College Entrance Exam, and passing the College English Test is the graduation requirement in most colleges.

[3]In particular, there are only limited TOEFL and GRE tests available in China, hence it is generally unlikely for Chinese students to make rush decisions to pursue graduate education abroad.

not required to use the English name outside the classroom. Researchers have explored various types of determinants of English-name usage, such as education (Fryer and Levitt, 2004), country and culture of origin (Abramitzky et al., 2014), and demographic characteristics (Sue and Telles, 2007). These are all important factors to be controlled.

However, social scientists have somewhat neglected another crucial factor related to name usage: the linguistic factor. A Chinese student would be more likely to use the English name if his Chinese name is difficult to pronounce in English, and English-name usage provides a solution to the pronunciation issue. This linguistic factor thus creates a natural experiment on English-name usage. In Appendix C of this dissertation, following the discussions in Chapter 3, I have introduced the criteria of the "pronunciation difficulty" based on linguistic properties of English and Chinese. Because most colleges (and even many middle and high schools) in China hire native speakers to teach English, Chinese students are exposed to English speakers early and find that pronunciation issues cause inconvenience and discomfort for both English and Chinese speakers. Therefore, for Chinese college students who plan to pursue graduate education abroad, they are able to realize that English-name usage could be a solution to the pronunciation issue even long before moving to the U.S.

A special feature of the pronunciation difficulty is that it is arguably exogenous in the sense that it is not associated with any personal or regional characteristics. This is due to the huge "linguistic difference" between two languages (Crowley and Bowern, 2010), and the pronunciation difficulty of a Chinese character in English does not have any implication in the context of Chinese. Moreover, the sample of this paper comprises students of the cohort born in the late

1980s, and most of the parents were born in the 1960s and had limited knowledge of English when their children were born. Therefore, it is unlikely that parents cared about the pronunciation difficulty by native speakers of *English*. In Section F.4 and Appendix D of the dissertation (following the appendix of Chapter 3), I will show that difficult-to-pronounce names are randomly "distributed" in the sample of this paper, and even the general Chinese population.

F.3.2 Academic Outcomes: School Tiers

This paper studies academic outcomes by examining what schools students attend. One way to evaluate school choices is to focus on school rankings. However, as there are many ranking systems that rank schools based on different criteria, an outcome variable constructed based on the exact rank is unlikely to be robust. Another way is to split schools into tiers, which constructs academic outcomes in a coarse manner. In this paper I categorize U.S. graduate schools in the sample into three tiers.

For U.S. graduate schools, I include universities in top 10 of the "US News Best Global University Rankings" and all other Ivy League schools in the first tier. This tier thus contains Harvard University, Massachusetts Institute of Technology, University of California-Berkeley, Stanford University, California Institute of Technology, University of California-Los Angeles, University of Chicago, Yale University, Columbia University, University of Pennsylvania, Cornell University, Brown University, and Dartmouth College. The second tier contains all other schools in the Association of American Universities (AAU), which comprises 62 leading universities in the U.S. and Canada. Note that all tier 1 schools

are AAU members. Finally, tier 3 schools include all other U.S. universities.

Undergraduate education is a major explanatory variable of graduate school outcomes. In this paper, I similarly split Chinese colleges into three tiers. Specifically, tier 1 colleges include all members of the C9 League[4], including Peking University, Tsinghua University, University of Science and Technology of China, Fudan University, Nanjing University, Shanghai Jiao Tong University, Zhejiang University, Harbin Institute of Technology, and Xian Jiao Tong University. Tier 2 colleges include universities sponsored by "Project 985"[5] but are not C9 League members. Tier 3 colleges include all other Chinese universities in the sample.

F.3.3 English Names and School Choices

As discussed earlier, English-name usage is correlated with immigrants' socioeconomic outcomes in general. English-name usage can similarly affect migrant students' academic outcomes through several channels. First, migrant students might rely on local names to avoid name-based discrimination in school. This explains the positive effect of English-name usage on subjective well-being and even test scores, but is not clearly concerning school choices. Second, English-name usage might be positively correlated with language or non-cognitive skills, and thus English-name users have better academic outcomes, such as test scores or educational attainment. However, as I focus only on advanced degree holders, skill disparities among these students should be minimal. That said, unlike test scores or educational attainment that are basically determined by

[4]Equivalent to the Ivy League in the U.S. and the *Exzellenzinitiative* in Germany, Chinas C9 League comprises nine most renowned universities in Mainland China.

[5]Project 985 is an official project initiated by national and local governments that allocate funding to 39 reputable research universities in Mainland China after careful evaluations on research and teaching quality.

skills, school choices can be affected by various factors other than skills.

This paper focuses on a crucial feature of school choice: a school is actually a combination of the school itself, and its locale. In other words, students make school choices based on both academic pursuits and locational preferences. Hence, even if English-name usage needs not to be related to skills, it can still lead to the choices of specific schools through its correlation with students' locational considerations.

When making locational choices, immigrants generally prefer areas with familiar demographic environments[6]: immigrants usually first choose ethnic enclaves (e.g., Bartel, 1989; Altonji and Card, 1991), and then areas with the large racial population sharing the low dissimilarity index with them[7]. However, cultural assimilation reduces the tendency of following such patterns. For example, measured by English proficiency, cultural assimilation decreases ethnic enclave residence (Bleakley and Chin, 2010). If English-name usage reflects efforts for cultural assimilation and similarly affects locational preferences, a Chinese student with English-name usage should be more willing to accept offers from top-tier schools in traditionally "less-chosen" areas, and the school choice is thus influenced.

[6]Also, see the general discussion of immigrants' geographic preferences: e.g., Farley and Haaga, 2000; Scott et al., 2005; Lymperopoulou, 2013; Sinha and Cropper, 2013.

[7]The dissimilarity index describes the degree of residential segregation between two ethnic groups.

F.4 Data and Empirical Strategies

In this section I introduce data and methods. I will first discuss the online social networking data set used in this paper. I then focus on the empirical strategies for identifying the effect of English-name usage on outcomes of graduate school admissions among Chinese students in the sample.

F.4.1 Data

In this paper, I use data retrieved from Renren, a Facebook-type social networking site founded in 2005. As Facebook is blocked, Renren is popular among college students in China and serves as Facebook's substitute[8]. Indeed, most college students in China have Renren accounts[9], and the selective registration of Renren should be a minor issue. I can also control for networking characteristics that reflect the popularity on Renren and the frequency of networking usage. Similar to Facebook data (e.g., Wimmer and Lewis, 2010), Renren provides users' biographical and educational information. A special advantage of the Renren sample is that the website gives users the option to add the English name following the Chinese name as the "suffix", and based on this I am able to measure English-name usage.

The Renren policy ensures that the Chinese name, English-name usage, and school attendance are all authentic information. At the time of data collection,

[8]Unlike Facebook, however, Renren is popular *only* among students, as its registration was only open for college students in China for a long time.

[9]Online social networking is widely recognized to be very popular among college students (Tella, 2014). Similarly, Renren is popular among college students in China. I conduct a simple test on this argument by observing Tsinghua students' Renren accounts. Tsinghua is one of the few schools that publicly release the list of enrolled students. I find that more than 90% of Tsinghua students have accounts on Renren.

the Renren policy does not allow users to change their usernames, and usernames used for registration are checked and approved by website administrators[10]. Hence users add English names no later than they update their graduate school information. Similar to Facebook, an .edu email address is needed for verifying the updated school information.

I restrict the sample to graduate students who arrive in the U.S. *straight after* earning bachelor's degrees in China. As discussed in Section 2, there is a clear trade-off pattern of English-name usage among these students, which leads to sufficient variation in English-name usage in the sample. Students who do not plan to receive education abroad also do not expect to be highly exposed to native speakers of English, and thus find it unnecessary to become English-name users in China.

A major concern about the measurement of English-name usage is that actual English-name usage might be mis-measured based only on online observations. For example, a student who shows the English name online might not use it in real life. In contrast, a student might still be the English-name user in real life even if he does not show the English name online, or shows an unusual name[11] that are not identified as a name. These might cause measurement error for the identification of the effect of English-name usage.

[10]On Renren, the national ID card or a Chinese phone number are needed for registration.

[11]Technically, Renren users can add any English word as the suffix that follows the Chinese names. In some cases, non-name words can be easily identified (such as *Mathematics*), but misidentification of English-name usage might occur in more ambiguous cases, as people from non-Anglophone countries are likely to adopt and use non-mainstream names (Edwards and Caballero, 2008) that are not in the dictionary of Anglicized given names widely used by native speakers of English.

F.4.2 Descriptive Statistics

I now describe the sample. Table F.1 presents the summary of independent variables. The first panel reports three basic variables. For the variable of the main interest—English-name usage—I find that 13.3% of students are English-name users on Renren. Nearly half of all students are male. The average year since entering college is approximately 8.7 years. Due to the data limitation I cannot observe age in the sample. However, this variable should serve as a good proxy for age, as I only focus on students who start graduate education straight after completing the undergraduate program. The concentration of the year of birth is around 1988.

The second panel describes school tiers. In the sample, 20% of students graduated from tier 1 Chinese colleges, and 27% from tier 2 colleges. Slightly more than half of students graduated from tier 3 colleges in China.

The last panel examines local demographic and socioeconomic characteristics of pre-migration cities. Over 30% and 40% of all students in the sample are originally from East Coast and Central North of China, respectively. These two areas are indeed the most populated regions in China. 7.2% of students in the sample are from Northeast, 10% are from Central South, and 7.5% are from provinces in Western China. Approximately 30% of all students are from coastal cities in China. In the city of origin, on average, the GDP per capita is nearly 100,000 Chinese Yuan, and the average human development index is 0.768. The average urban population is 13.5 million, the average area of the city is 325 square miles, and the average density is about 40,000 per square mile.

Starting from Table F.2, I turn to focus on variables related to the U.S., i.e.,

Table F.1: Summary of Independent Variables

	Mean	Std. dev.
Basic Variables:		
English-name usage	0.133	(0.340)
Male dummy	0.489	(0.500)
Year since entering college	8.684	(1.826)
College-Tier Variables:		
Tier 1 Chinese college	0.205	(0.403)
Tier 2 Chinese college	0.270	(0.444)
Tier 3 Chinese college	0.525	(0.499)
Pre-Migration Variables:		
Region 1: East Coast	0.314	(0.464)
Region 2: Central North	0.435	(0.496)
Region 3: Northeast	0.075	(0.263)
Region 4: Central South	0.104	(0.305)
Region 5: West	0.072	(0.258)
Coastal city	0.287	(0.452)
GDP per capita (CNY)	97166.910	(20704.800)
Human development index	0.768	(0.566)
Population (urban)	1.350e+07	(6.961e+06)
Area (urban, sq mi)	325.185	(141.155)
Density (urban)	40933.620	(10951.830)
Observations	7,287	

school variables and local demographic characteristics of the locales of the graduate schools. 14% of all students in the sample attend tier 1 U.S. schools and nearly half of all students attend tier 2 U.S. schools. Schools in the first two tiers consist of Association of American Universities (AAU) schools, and 63.8% of students in the sample attend graduate schools that are members of AAU. 44.8% of students attend private schools in the U.S.

Table F.2: Summary of Dependent Variables: Graduate School Tiers

	Mean	Std. dev.
Tier 1 U.S. graduate school	0.140	(0.347)
Tier 2 U.S. graduate school	0.498	(0.500)
Tier 3 U.S. graduate school	0.362	(0.481)
AAU schools	0.638	(0.481)
Private school	0.448	(0.497)
Observations	7,287	

Table F.3: Summary of Local Demographic Characteristics in the U.S.

	Mean	Std. dev.
% White residents	0.615	(0.147)
% Asian residents	0.090	(0.064)
% Black residents	0.207	(0.174)
Observations	7,287	

On average, in the local area of the graduate school attended, the percentage of White residents is 61.5%. The percentage of Asian residents is 9% and the percentage of Black residents is 20.7%. In general, the racial composition is different from the national average, mainly because that the percentages are "geographically weighted", in the sense that larger cities host more schools and also have more minorities (including international students from China), compared with the national average.

In Table F.4 I examine the interaction terms between school and local demographic variables. In the main empirical analysis, I focus on graduate schools located in areas with *large* (or *small*) White, Asian, and Black populations. Specifically, I consider an area with *large* White populations if more than 60% of local residents are White; for Asian and Black populations this percentage is 10%

Table F.4: Summary of Dependent Variables: Interaction Terms between School Characteristics and Local Demographic Characteristics in the U.S.

School term		Geo. term	# of students	Mean	Std. dev.
Tier 1 U.S. school	×	% White > 60%	547	0.075	(0.263)
Tier 1 U.S. school	×	% White < 50%	472	0.065	(0.246)
AAU school	×	% White < 50%	1,108	0.152	(0.359)
Tier 1 U.S. school	×	% Asian > 10%	367	0.050	(0.219)
Tier 1 U.S. school	×	% Asian < 5%	221	0.030	(0.171)
AAU school	×	% Asian < 5%	1,118	0.153	(0.360)
Tier 1 U.S. school	×	% Black > 30%	418	0.057	(0.233)
AAU school	×	% Black > 30%	1,054	0.145	(0.352)
Tier 1 U.S. school	×	% Black < 10%	396	0.054	(0.227)
AAU school	×	% Black < 10%	2059	0.283	(0.420)
Total observations			7,287		

and 30%, respectively. On the other hand, I consider an area with *small* White populations if less than half of local residents are White; for Asian and Black populations this percentage is 5% and 10%, respectively. In the latter analysis I will test the sensitivity by using other thresholds of percentages.

Table F.4 shows that 7.5% of all students in the sample attend tier 1 U.S. schools in areas with large White populations. On the other hand, 6.5% of students attend tier 1 schools in areas with small White populations, and 15.2% of students in the sample attend AAU (tier 1 & 2) schools in such areas. Subsequently, 5% of students attend tier 1 schools in areas with large Asian populations. Only 3% of students attend tier 1 U.S. schools in areas with small Asian populations, while much more (15% of all) students in the sample enter AAU schools in such areas. Finally, I examine interactions between school tiers and Black populations. 5.7% of students in the sample enter tier 1 U.S. schools in areas with large Black populations, and 14.5% of students enter AAU schools

in such areas. On the other hand, 5.4% of students enter tier 1 U.S. schools in areas with small Black populations, and 28.3% of students enter AAU schools in such areas. Table F.4 indicates the heterogeneous geographic distribution of Chinese graduate students in the U.S., even after taking the school tier into consideration.

F.4.3 Empirical Strategies

Let T_i be individual i's school or school-location choice, and E_i be the indicator of i's English-name usage shown on Renren. I start with the baseline estimate of the effect of English-name usage on graduate school choice using OLS:

$$T_i = \beta_0 + \beta_1 E_i + \mathbf{X}_i \beta_2 + \varepsilon_i \tag{F.1}$$

where $\mathbf{X}_i$ is the vector of covariates and ε_i is the error term. In this model, β_1 is the effect of English-name usage on T_i. However, β_1 is possibly biased due to the endogeneity of E_i. The major concern of estimating the effect of cultural assimilation is reversal causality, i.e., the assimilation skill or behavior might be the consequence instead of cause of the social outcome (see, e.g., Bleakley and Chin, 2010). In this paper, however, the issue of reversal causality should be minor as students are allowed to update school information but not their usernames. The only exceptions are that the graduate school choice is pre-determined—through, e.g., the campus visit and oral agreement—before the Renren account is registered, which cannot be captured by the data set.

Another issue is that E_i is likely to be mis-measured. For example, if a student uses an English name in reality but does not show the name on Renren,

then his English-name usage is not captured in this online sample. In contrast, it is also possible that a student shows an English name online but does not use it in reality, i.e., he is incorrectly identified as an English-name user. Furthermore, there are also many ambiguous cases in which it is difficult to determine whether the English word presented on Renren is a name or not[12].

The above issues imply that E_i is endogenous. The standard solution to the endogeneity problem is to use an exogenous variable to instrument for E_i. In this paper, I use the difficulty of pronouncing the Chinese name in English, classified based on linguistic properties of two languages, to instrument for English-name usage. Let P_i be the individual-level pronunciation difficulty indicator for i. The identification is based on the assumption that P_i predicts E_i and influences the outcome, T_i, only through its effect on E_i. This instrumental variable (IV) approach estimates the effect of English-name usage using two-stage least squares (TSLS), led by the following first-stage regression:

$$E_i = \alpha_0 + \alpha_1 P_i + \mathbf{X}_i \alpha_2 + \epsilon_i \tag{F.2}$$

F.4.4 The Validity of the IV

I now discuss the validity of the IV. To ensure its validity, the pronunciation difficulty indicator should first be a good predictor of English-name usage. In Column 1, Table F.5, I run the regression of English-name usage only on the pronunciation difficulty dummy. The correlation is significant and strong; the result shows that an individual whose Chinese name is difficult to pronounce

[12]Social scientists have long observed that immigrants from non-Anglophone countries might choose non-mainstream Anglicized names that are generally not in the name dictionary (Edwards and Caballero, 2008). This is also true for English names among native speakers of Chinese.

in English is approximately 12.5% more likely to use the English name as presented on Renren.

Table F.5: First-Stage Regressions

	English-name usage				
	(1)	(2)	(3)	(4)	(5)
Pronunciation difficulty	0.125***	0.123***	0.123***	0.123***	0.122***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Individual demographics	No	Yes	Yes	Yes	Yes
School tier FE	No	No	Yes	Yes	No
Pre-migration geographics	No	No	No	Yes	Yes
School FE	No	No	No	No	Yes
F-statistics	246.92	243.34	244.62	243.67	230.56

Observations: 7,287. Standard errors are in parentheses. *: $p < .1$; **: $p < .05$; ***: $p < .01$.

In the following columns I add other control variables and rerun the first-stage regression. In Column 2 I only include individual demographic variables and find that the extent of the relationship between the pronunciation difficulty and English-name usage only becomes slightly smaller. Including school tier fixed effects and pre-migration geographic variables in Column 3 and 4, the extent of the first-stage relationship remains stable. In Column 5 I use Chinese college fixed effects instead of school tier fixed effects as the covariates, and the magnitude of the influence of pronunciation difficulty on English-name usage does not change. In general, I find that students with difficult-to-pronounce names are indeed significantly more likely to be English-name users.

In Table F.6 I conduct balancing tests to check the difference in observable characteristics between two groups of students split by pronunciation difficul-

ty. As discussed earlier, because most parents had limited English proficiency when naming decisions were made, the pronunciation difficulty should not be associated with any individual and regional characteristics. Indeed, I find no systematic difference in gender and the tier of the Chinese college attended between students with and without difficult-to-pronounce names. While the association between the year since entering college and the pronunciation difficulty is significant, its magnitude is subtle at best.

Table F.6: Checking on Systematic Differences

	w/o difficult-to-pronounce names	w/ difficult-to-pronounce names	*p*-value
Male	0.486 (0.500)	0.493 (0.500)	n.s.
Year since entering college	8.732 (1.822)	8.619 (1.828)	**
Category 1 Chinese college dummy	0.203 (0.402)	0.209 (0.407)	n.s.
Category 2 Chinese college dummy	0.271 (0.445)	0.267 (0.443)	n.s.
Category 3 Chinese college dummy	0.527 (0.445)	0.523 (0.443)	n.s.
East Coast	0.317 (0.465)	0.309 (0.462)	n.s.
Central North	0.439 (0.496)	0.431 (0.495)	n.s.
Northeast	0.069 (0.254)	0.082 (0.275)	*
Central South	0.104 (0.305)	0.105 (0.306)	n.s.
West	0.071 (0.245)	0.074 (0.261)	n.s.
GDP per capita (log)	11.458 (0.264)	11.447 (0.272)	n.s.
Human development index	0.768 (0.055)	0.766 (0.066)	n.s.
Population (log)	16.244 (0.678)	16.222 (0.676)	n.s.
Area	848.899 (366.487)	833.092 (364.220)	n.s.
Density (log)	9.634 (0.267)	9.631 (0.268)	n.s.
Observations	4,210	3,077	

Standard deviations are in parentheses.

Unpaired t tests are employed. n.s.: $p \geq .05$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In the rest of the table I examine whether the pronunciation difficulty is as-

sociated with specific regions or local socioeconomic and demographic characteristics. In general, I find that there is almost no significant difference in any regional characteristics between two groups of students. The pronunciation difficulty is neither associated with certain regions[13] nor city-level socioeconomic and demographic characteristics. Based on the above balancing tests for observable variables, I find that it is arguably random whether a Chinese name is identified as "difficult to pronounce". Moreover, in Appendix D of the dissertation I show that the percentage of difficult-to-pronounce names in this sample appears to be close to that in external samples that contain non-migrant students, indicating that the pronunciation difficulty is even not associated with students' migration decisions.

F.5 Empirical Analysis: School Choices and School-Location Outcomes

This section reports the empirical findings of this paper. I first present main results, and then discuss the findings and conduct several tests to check the robustness of the main results.

F.5.1 Main Results

Table F.7 reports the main results of this paper. In this table, I regress the academic outcome, measured by the tier of the graduate school attended, on

[13]The only exception is that there are slightly more students from Northeast China who have difficult-to-pronounce Chinese names, but this might be simply due to small-sample bias as the percentage of college students from this region is relatively low.

English-name usage and other covariates. In Column 1 I regress tier 1 U.S. school attendance on English-name usage and individual characteristics, and find no statistically significant relationship between English-name usage and tier 1 school attendance. In Column 2 I use the pronunciation difficulty indicator to instrument for English-name usage and estimate the effect of English-name usage using TSLS. Similarly, I observe no significant effect of English-name usage. I repeat the exercise in Column 3 and 4 by including pre-migration local socioeconomic and demographic characteristics as covariates; still, both the OLS and IV estimation suggest no significant effect of English-name usage on tier 1 school attendance.

Table F.7: English-Name Usage and Academic Outcomes

	Tier 1 U.S. graduate school attendance				AAU (tier 1 & tier 2) school attendance			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
English-name usage	−0.012	0.028	−0.007	0.040	0.031†	0.011	0.037**	0.041
	(0.011)	(0.063)	(0.011)	(0.062)	(0.016)	(0.090)	(0.016)	(0.089)
Male	−0.006	−0.005	−0.005	−0.003	−0.017	−0.017	−0.013	−0.013
	(0.008)	(0.008)	(0.008)	(0.008)	(0.011)	(0.011)	(0.010)	(0.011)
Year since entering	−0.011***	−0.011***	−0.011***	−0.010***	−0.004	−0.005	−0.004	−0.004
college	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003	(0.003)	(0.004)
Tier 1 college	0.312***	0.313***	0.305***	0.307***	0.319***	0.318***	0.297***	0.297***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.014)	(0.015)	(0.015)	(0.015)
Tier 2 college	0.033***	0.034***	0.032***	0.032***	0.106***	0.116***	0.114***	0.114***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.013)	(0.013)	(0.013)	(0.013)
Local covariates	No	No	Yes	Yes	No	No	Yes	Yes

Observations: 7,287. Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In the rest of the table I rerun the above four regressions, but now using AAU school (tier 1 & 2 schools) attendance as the dependent variable. In Column 5

and 7, the OLS estimates imply that English-name usage is associated with AAU school attendance. However, after instrumenting English-name usage using the pronunciation difficulty indicator, I again find no significant effect of English-name usage on AAU attendance. In general, Table F.7 implies that English-name usage does not significantly affect the tier of the U.S. graduate school attended.

In Table F.8 I examine the heterogeneous effect of English-name usage by gender. Social scientists have long observed the heterogeneity in education by gender (e.g., Buchmann et al., 2008) and the determinants of educational outcomes might have different effects on male and female students (e.g., Autor et al., 2016). In the first row of the table I report regression results in the sample of male students. Similar to the overall effect, the effect of English-name usage on male students' school tiers is minor. However, English-name usage does affect (significant at the 0.1 level) the choices of top-tier schools among female students. Although there is no overall effect of English-name usage on students' school choices, Table F.8 still shows some evidence of heterogeneous effects by gender.

Table F.8: English-Name Usage and Academic Outcomes by Gender

		(1)	(2)	(3)	(4)	
Gender	Model	Tier 1 U.S. graduate school	Tier 1 U.S. graduate school	AAU (tier 1 & 2) graduate school	AAU (tier 1 & 2) graduate school	Observations
Male	IV	−0.052	−0.059	−0.020	−0.032	3,557
		(0.107)	(0.101)	(0.145)	(0.139)	
Female	IV	0.132	0.134†	0.064	0.093	3,730
		(0.085)	(0.079)	(0.121)	(0.115)	
Local covariates		No	Yes	No	Yes	

Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

In Table F.9 I turn to study joint school-location outcomes. I study locational

choices by focusing on local demographic characteristics. Immigrants traditionally prefer ethnic enclave residence (Bartel, 1989), and then in areas with larger racial populations that have low dissimilarity indexes with Asian. In the context of this paper, Asian and White have a fairly low index of dissimilarity (Iceland et al., 2014).

In the first panel I study interactions between tier 1 school attendance and local demographics. While OLS regressions show no effect of English-name usage on the interaction between school and locational choices, IV estimates imply that Chinese students with English-name usage are more likely to attend tier 1 schools in one of the following types of areas, including areas with (a) small White populations; (b) small Asian populations; and (c) large Black populations. As discussed earlier, these areas are traditionally "less chosen" by Chinese migrants. Hence, Table F.9 implies that English-name usage is correlated with school choices when locational characteristics are taken into consideration, and thus English-name usage affects joint school-location choices.

However, Table F.9 shows no effect on top-tier school attendance in areas with (a) large White populations; (b) large Asian populations; and (c) small Black populations. These areas are considered to be traditionally popular as Asian immigrants prefer ethnic enclave residence and areas with "residentially familiar" populations. In this paper, however, English-name usage does not affect tier 1 school attendance in such areas. In addition, the second panel implies that English-name usage does not affect AAU school attendance in any areas shown in the table. These suggest that although English-name usage affects Chinese graduate students' joint school-location choices, its effect is significant only when students consider top-tier schools in traditionally "less-chosen" ar-

Table F.9: English-Name Usage, Geographic Characteristics, and School Tier

		(1)	(2)	(3)	(4)	(5)	(6)
Model	Interaction:	% White > .6	% White < .5	% Asian > .1	% Asian < .05	% Black > .3	Black < .1
OLS	Tier 1 U.S.	−0.008	0.001	−0.004	0.004	−0.002	−0.007
	school	(0.009)	(0.008)	(0.005)	(0.006)	(0.008)	(0.008)
IV	Tier 1 U.S.	−0.045	0.085†	−0.035	0.154***	0.098*	−0.041
	school	(0.049)	(0.047)	(0.039)	(0.034)	(0.045)	(0.043)
OLS	AAU school	0.033*	−0.006	0.018	−0.014	−0.013	0.002
	(tier 1 & 2)	(0.017)	(0.008)	(0.015)	(0.013)	(0.012)	(0.016)
IV	AAU school	0.094	−0.010	0.253**	−0.046	−0.038	0.059
	(tier 1 & 2)	(0.094)	(0.012)	(0.086)	(0.069)	(0.067)	(0.087)

Only the coefficient of English-name usage is presented in this table. However, all covariates are included.

Observations: 7,287. Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

eas in the U.S.

F.5.2 Discussions: Main Results

The above results imply that the overall effect of English-name usage on graduate school admissions is minor. As students who plan to pursue graduate studies have high levels of human capital accumulation, English-name usage might not lead to significant better skills and have little or no value added. Even if English-name usage does positively affect skills, these might not translate into better admission results[14].

The above findings also imply that compared with English-name users, students without English-name usage have similar admission outcomes (as shown

[14]Indeed, researchers find that even the direct improvement in assimilation-related skills (e.g., language) does not affect minorities' educational outcomes (Chin et al., 2007).

in Table F.7) even if they are unwilling to attend tier 1 schools in traditionally "less-chosen" areas. In Table F.10 I examine this result by regressing tier 1 school attendance on English-name usage conditional on specific local demographic characteristics in the U.S.

Table F.10: English-Name Usage and School Tier Conditional on Local Demographics

		(1)	(2)	(3)	(4)	(5)	(6)
Model	Conditional on:	% White > .6	% White < .5	% Asian > .1	% Asian < .05	% Black > .3	Black < .1
IV	Tier 1 U.S.	−0.074	0.425*	−0.223*	0.775***	0.543**	−0.063
	school	(0.070)	(0.192)	(0.101)	(0.188)	(0.208)	(0.092)
	Observations	4,408	1,722	2,119	1,949	1,779	2,022

Only the coefficient of English-name usage is presented in this table. However, all covariates are included.

Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Regression results in Column 2, 4, and 5 show that conditional on choosing "less-chosen" areas—with small Asian or White populations, or large Black populations—English-name usage is positively correlated with the school tier, which is consistent with earlier findings. On the other hand, English-name usage is negatively correlated with tier 1 school attendance in areas with large Asian populations, and has no effect in areas with large White or small Black populations. These imply that students without English-name usage are able to attend schools at the similar level with English-name users because they are able to receive similar or even better offers from school located in traditionally "popular" areas, and thus in general, the effect of English-name usage on the tier of the school attended is insignificant.

F.5.3 Additional Tests: Sensitivity

One concern about the main results is that the percentages of ethnic populations involved in Table F.9 and F.10 are sensitive. In Table F.11, I conduct several sensitivity tests to check the robustness of the main results.

Table F.11: English-Name Usage, Local Demographic Characteristics, and School Choices

		(1)	(2)	(3)	(4)	(5)	(6)
Model	Interaction:	% White < .6	% Asian < .03	% Asian < .08	% Asian < .1	% Black > .2	% Black > .4
IV	Tier 1 U.S.	0.085†	0.154***	0.085†	0.075	0.085†	0.154**
	school	(0.047)	(0.034)	(0.049)	(0.055)	(0.047)	(0.034)
IV	AAU school	−0.053	0.075	−0.210*	−0.212*	−0.112	−0.017
	(tier 1 & 2)	(0.080)	(0.051)	(0.088)	(0.093)	(0.081)	(0.062)

Only the coefficient of English-name usage is presented in this table. However, all covariates are included.

Observations: 7,287. Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

Column 1 examines tier 1 school attendance in areas with small White populations, and now using 60% as the threshold. This percentage is higher than the 50% threshold used in earlier regressions and is close to the national average of the percentage of the (non-Hispanic) White. Still, English-name usage positively affects tier 1 school attendance in such areas, although there is again no effect on AAU school attendance.

I turn to focus on local Asian populations from Column 2 to 4. Asian Americans comprise 4.8% of the population in the U.S., which is close to the percentage in Table F.9 and F.10. Adjusting the threshold of the low Asian population to either 3% or 8% does not alter the qualitative pattern of the effect of English-name usage on the interaction between school and locational outcome. However, the effect of English-name usage on tier 1 school attendance becomes

insignificant when using 10% of local residents to define the *large* Asian population.

In Column 5 and 6 I focus on local Black populations. In the previous table, the threshold of the *large* Black population is 30%, and in these columns I use 20% and 40% as the threshold. Both columns show the robustness of the main result: English-name usage is positively correlated with tier 1 school attendance in areas with large Black populations in both Column 5 and 6, regardless of the threshold. This is consistent with the qualitative pattern observed earlier.

F.5.4 Additional Tests: Heterogeneous Effects by Gender

I conclude the empirical section by revisiting heterogeneous effects of English-name usage by gender. In earlier regressions I have shown the effect of English-name usage on the tier of the graduate school attended differs among male and female students, and in Table F.12 I examine its effect on school-location choices in two sub-samples.

In Table F.9, the general results indicate that English-name usage is positively correlated with tier 1 school attendance in traditionally "less-chosen" areas, where Asian or White populations are relatively small, or Black populations are relatively large. Table F.12 further shows the heterogeneous effect of English-name usage on joint school-location choices by gender among Chinese students, and the effect of English-name usage on school-location outcomes is significant mainly among female students.

Specifically, English-name usage does improves the tendency of tier 1 school

Table F.12: English-Name Usage and School-Location Choices by Gender

		(1)	(2)	(3)	
Gender	Model	(Tier 1 school) × % White < .5	(Tier 1 school) × % Asian < .05	(Tier 1 school) × % Black > .3	Observations
Male	IV	0.033	0.156**	0.051	3,557
		(0.073)	(0.052)	(0.068)	
Female	IV	0.138*	0.157**	0.147*	3,730
		(0.061)	(0.045)	(0.059)	

Only the coefficient of English-name usage is presented.

However, all covariates are included.

Standard errors are in parentheses. †: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$.

attendance in areas with small Asian populations among both female and male students, and I observe similar effect sizes in both sub-samples (around 16%). However, English-name usage appears to be unrelated to male students' tier 1 school attendance in other two types of "less-chosen" areas, while significantly improve the tendency of tier 1 school attendance in areas with small White populations or large Black populations among female students. In general, although overall English-name usage affects joint school-location outcomes, its effect is highly heterogeneous and, in most cases, only female students are affected.

F.6 Conclusion

This paper examines the effect of efforts for cultural assimilation on U.S. school choices among graduate school applicants from China. Specifically, in this paper I use English-name usage to measure efforts for cultural assimilation and examine its effect on the tier of the U.S. graduate school attended.

In this paper, I use online social networking data retrieved from Renren and focus on students who receive undergraduate education in China and graduate education in the U.S. English-name usage can be endogenous because (a) the tier of the school attended might reversely affect English-name usage, and (b) there are measurement issues concerning online English-name usage. I solve these problems by exploiting a linguistic natural experiment on English-name usage: a student is more likely to use the English name if his Chinese name is difficult to pronounce in English. The pronunciation difficulty indicator serves as the IV for English-name usage. Balancing tests imply that the "pronunciation difficulty" almost randomly exists among Chinese names in the sample.

I employ this method to estimate the effect of English-name usage on Chinese students' school choices. I find no overall effect of English-name usage on the tier of the graduate school attended, but the effect is heterogeneous: while English-name usage has no effect on male students' school tiers, it increases female students' top-tier school attendance. Moreover, English-name usage affects the interaction between the school tier and the local racial makeup. Because immigrants usually prefer areas with large populations of their own ethnic groups or large populations that share the low dissimilarity index (e.g., Massey and Denton, 1985; Bartel, 1989), Chinese migrants traditionally do not choose areas with relatively small Asian or White populations, or large Black populations. I find that English-name usage increases top-tier school attendance in these traditionally "less-chosen" areas. Again, English-name usage mainly affects female students' school-location choices and, in most cases, have insignificant impacts on male students.

By focusing on the specific example of English-name usage, this paper points

out the possible role efforts for cultural assimilation in choosing schools and destination areas among foreign-born advanced degree holders. Specifically, this paper highlights the importance of considering immigrant students' school choices as joint choices along multiple dimensions, namely, academic pursuits and locational preferences.

APPENDIX G

APPENDIX FOR CHAPTER 4, PART A: THE CONSTRUCTION OF THE IV

In this appendix, I introduce the construction of the achievement variable, which is used for instrumenting for the network size. As introduced in Section 4.3, there has been no barrier for English teams to purchase French players after the "Bosman Ruling" came into effect. The English Premier League is richer and more prestigious than the French league, thus French players have the professional incentive to migrate; besides, English teams are keen to buy foreigners from EU states that produce better players, such as France. Hence the achievement of the France national team directly affects the influx of French players. To get started, I need to find a proxy for the achievement of France.

I first obtain France's "relative ranking" r_k at the k-th tournament for national teams:

$$r_k = 1 - \frac{R_k - 1}{N_k} \tag{G.1}$$

i.e., r_k is the proportion of teams that are outperformed by France in this tournament. R_k is the "absolute" ranking of the France national team (e.g., $R_k = 1$ if France is the winner, $R_k = 2$ if France is the runner-up, etc.), and N_k is the number of participants of the tournament. To enhance the robustness and take timing into account, I calculate France's average ranking in the past two tournaments, i.e.,

$$A(y_k) = \frac{r_k + r_{k-1}}{2} \tag{G.2}$$

where y_k are years no later than the k-th and before the $(k+1)$ tournament, and $A(y_k)$ is the achievement variable for these years. After obtaining the achievement of France in each year, I can define the achievement variable for

the network: for a French player in England, this achievement variable is the average achievement of France upon each of his compatriot teammate's arrival. I report the equation to average France's achievement upon arrival of teammates in Section 4.5, and the related first-stage regressions in Section 4.6.

APPENDIX H

APPENDIX FOR CHAPTER 4, PART B: OTHER NOTES ON THE VALIDITY OF THE IV

In this appendix I discuss several additional tests on the validity of the constructed achievement variable as the IV. The basic idea of using the achievement of the France national team as the IV is that there is only a small fraction of French players who are able to play for their nation. But in addition, the validity of the IV will be clearer if the changing pattern of the skill distribution of French players who do not represent their national team does not significantly affect quality of players migrating to England.

Table H.1: French Youth Team in U-20 World Cup and (U-23) Olympic Games

Year	Achievement	# of teams	Year	Achievement	# of teams
1987	not qualified	16	2000	not qualified	16
1988	not qualified	16	2001	6th place	24
1989	not qualified	16	2003	not qualified	24
1991	not qualified	16	2004	not qualified	16
1992	not qualified	16	2005	not qualified	24
1993	not qualified	16	2007	not qualified	24
1995	not qualified	16	2008	not qualified	24
1996	5th place	16	2009	not qualified	24
1997	7th place	16	2011	4th place	24
1999	not qualified	24	2012	4th place	24

Tournaments held in 1988, 1992, 1996, 2000, 2004, 2008, 2012 are Olympic Games.

In Table H.1 I present France's performance in U-20 World Cups and U-23 Olympic Games from 1987 to 2012. The national team only contains a small number of players, but there are much more players who have ever been enrolled in the youth team, and many of them finally migrated to England. The trend of France's achievement in tournaments for youth national teams are fairly stable, which reflects that the general skill distribution among French players is unlikely to change radically, and while not reported, I also find no correlation between French players' league appearances in England and the achievement variable constructed by the achievement of French youth national team.

Table H.2: French Youth Team in European U-21 Championship

Year	Achievement	# of teams	Year	Achievement	# of teams
1986	7th place	8	2000	not qualified	8
1988	winner (1st place)	8	2002	runner-up (2nd place)	8
1990	not qualified	8	2004	not qualified	8
1992	not qualified	8	2006	3rd place	8
1994	4th place	8	2007	not qualified	8
1996	3rd place	8	2009	not qualified	8
1998	not qualified	8	2011	not qualified	8

In Table H.2 I present France's achievement in European U-21 Championships. I repeat the exercise and construct a variable based on the achievement of the France youth national team in European tournaments. Still, I see from Table H.2 that there is no much variation in France's achievement, and the achievement variable constructed based on data in this table is not a robust predictor of French players' league appearances in England.

Table H.3: Regression of League Appearances on Year of Arrival Fixed Effects

Year	Coef.	Std. err.	p-value	Year	Coef.	Std. err.	p-value
1995	−19.230	11.604	0.101	2004	−12.369	10.129	0.224
1996	−15.145	10.129	0.137	2005	−17.513	10.593	0.101
1997	−19.249	9.832	0.053	2006	−21.800	10.379	0.038
1998	−19.606	9.861	0.049	2007	−17.501	10.129	0.087
1999	−15.801	9.987	0.116	2008	−14.106	10.129	0.166
2000	−17.230	9.896	0.084	2009	−19.263	10.940	0.081
2001	−13.610	9.697	0.163	2010	−19.991	10.233	0.053
2002	−19.295	10.049	0.108	2011	−29.513	10.940	0.008
2003	− 19.653	9.832	0.048	Constant	34.180	9.475	< 0.001

In Table H.3 I construct a cross-sectional player-career sample and run the regression of average league appearances across the player's career in England on the dummy of his year of first arrival in England. However, the relationship between the arrival year and league appearances in England is mostly unclear, as reported in this table. In particular, players who arrived in years when France achieved high rankings (i.e., 1996-2001, 2006-2007) do not really earn significantly more league appearances in England. This indicates that although the achievement variable does predict the size of the ethnic network, I find no correlation between year of arrival fixed effects and average appearances in the English Premier League among French players in England.

BIBLIOGRAPHY

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2014. "Cultural Assimilation during the Age of Mass Migration." manuscript.

Adler, Patricia A., and Peter Adler. 1984. "The Carpool: A Socializing Adjunct to the Educational Experience." *Sociology of Education*, 57(4), 200 - 210.

Ahmed, Ali M., Mats Hammerstedt. 2008. "Discrimination in the Rental Housing Market: A Field Experiment on the Internet." *Journal of Urban Economics*, 64(2), 362 - 372.

Alba, Richard, and Victor Nee. 2005. *Remaking the American Mainstream: Assimilation and Contemporary Immigration*. Cambridge: Harvard University Press.

Altonji, Joseph G., and David Card. 1991. "The Effects of Immigration on the Labor Market Outcomes of Less-skilled Natives." In *Immigration, Trade, and the Labor Market*, eds., John Abowd and Richard Freeman. Chicago: University of Chicago Press.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, 91(434), 444 - 455.

Arai, Mahmood, and Peter Skogman Thoursie. 2009. "Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings." *Journal of Labor Economics*, 27(1), 127 - 147.

Aral, Sinan, Lev Muchnik, and Arun Sundararajan. 2009. "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Net-

works." *Proceedings of the National Academy of Sciences*, 106(51), 21544 - 21549.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2016. "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes." manuscript.

Bard, Erin. 1997. "Transit and Carpool Commuting and Household Vehicle Trip Making: Panel Data Analysis." *Transportation Research Record: Journal of the Transportation Research Board*, 1598, 25 - 31.

Bartel, Ann P. 1989. "Where Do the New U.S. Immigrants Live?" *Journal of Labor Economics*, 7(4), 371 - 391.

Bassetti, Benedetta. 2007. "Effects of Hanyu Pinyin on Pronunciation in Learners of Chinese as a Foreign Language." In *The Cognition, Learning and Teaching of Chinese Characters*, eds., Andreas Guder, Xin Jiang, and Lexin Wang. Beijing: Beijing Language and Culture University Press.

Battu, Harminder, McDonald Mwale, and Yves Zenou. 2007. "Oppositional Identities and the Labor Market." *Journal of Population Economics*, 20(3), 643 - 667.

Battu, Harminder, and Yves Zenou. 2010. "Oppositional Identities and Employment for Ethnic Minorities: Evidence from England." *Economic Journal*, 120(542), F52 - F71.

Becker, Gary S. 1965. "A Theory of the Allocation of Time." *Economic Journal*, 75(299), 493 - 511.

Beckhusen, Julia, Raymond J. G. M. Florax, Thomas de Graaff, Jacques Poot, and Brigitte Waldorf. 2013. "Living and Working in Ethnic Enclaves: English Language Proficiency of Immigrants in US Metropolitan Areas." *Papers in Regional Science*, 92(2), 305 - 328.

Bell, Brian, and Stephen Machin. 2013. "Immigrant Enclaves and Crime." *Journal of Regional Science*, 53(1), 118 - 141.

Belot, Michèle, and Sjef Ederveen. 2012. "Cultural Barriers in Migration between OECD Countries." *Journal of Population Economics*, 25(3), 1077 - 1105.

Bennett, Daniel, Chun-Fang Chiang, and Anup Malani. 2015. "Learning during a Crisis: The SARS Epidemic in Taiwan." *Journal of Development Economics*, 112, 1 - 18.

Berman, Eli, Kevin Lang, and Erez Siniver. 2003. "Language-Skill Complementarity: Returns to Immigrant Language Acquisition." *Labour Economics*, 10(3), 265 - 290.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4), 991 - 1013.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, 100(5), 992 - 1026.

Bisin, Eleonora Patacchini, Thierry Verdier, and Yves Zenou. 2011. "Ethnic Identity and Labour Market Outcomes of Immigrants in Europe." *Economic Policy*, 26(65), 57 - 92.

Blásquez, Maite, Carlos Llano, and Julian Moral. 2010. "Geographical Barriers to Employment for American-born and Immigrant Workers." *Urban Studies*, 47(8), 1663 - 1686.

Bleakley, Hoyt, and Aimie Chin. 2004. "Language Skills and Earnings: Evidence from Childhood Immigrants." *Review of Economics and Statistics*, 86(2), 481 - 496.

Bleakley, Hoyt, and Aimie Chin. 2010. "Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants.." *American Economic Journal: Applied Economics*, 2(1), 165 - 192.

Blumenberg, Evelyn, and Michael Smart. 2010. "Getting by with a Little Help from My Friends... and Family: Immigrants and Carpooling." *Transportation*, 37(3), 429 - 446.

Blumenberg, Evelyn, and Michael Smart. 2014. "Brother Can You Spare a Ride? Carpooling in Immigrant Neighbourhoods." *Urban Studies*, 51(9), 1871 - 1890.

Blumenstock, Joshua, Nathan Eagle, and Marcel Fafchamps. 2016. "Airtime Transfers and Mobile Communications: Evidence in the Aftermath of Natural Disasters." *Journal of Development Economics*, 120, 157 - 181.

Borgatti, Stephen P., and Rob Cross. 2003. "A Relational View of Information Seeking and Learning in Social Networks." *Management Science*, 49(4), 432 - 445.

Borjas, George J. 1987. "Self-Selection and the Earnings of Immigrants." *American Economic Review*, 77(4), 531 - 553.

Borjas, George J. 1994. "The Economics of Immigration." *Journal of Economic Literature*, 32(7), 1667 - 1717.

Borjas, George J. 1995. "Assimilation and Changes in Cohort Quality Revisited: What Happened to Immigrant Earnings in the 1980s?" *Journal of Labor Economics*, 13(2), 201 - 245.

Borjas, George J. 1998. "To Ghetto or Not to Ghetto: Ethnicity and Residential Segregation." *Journal of Urban Economics*, 44(2), 228 - 253.

Borjas, George J. 1999. *Heaven's Door: Immigration Policy and the American Economy*, Princeton, NJ: Princeton University Press.

Borjas, George J. 2015. "The Wage Impact of the Marielitos: A Reappraisal." NBER Working Paper No. 21588.

Borjas, George J., and Kirk B. Doran. 2012. "The Collapse of the Soviet Union and the Productivity of American Mathematicians." *Quarterly Journal of Economics*, 127(3), 1143 - 1203.

Borjas, George J., Kirk B. Doran, and Ying Shen. 2015. "Ethnic Complementarities after the Opening of China: How Chinese Graduate Students Affected the Productivity of Their Advisors." NBER Working Paper No. 21096.

Boyd, Monica. 1989. "Family and Personal Networks in International Migration: Recent Developments and New Agendas." *International Migration Review*, 23(3), 638 - 670.

Bound, John, Charles Brown, and Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, volume 5, ed. James J. Heckman and Edward E. Leamer. Amsterdam: Elsevier.

Brownstone, David, Arindam Ghosh, Thomas F. Golob, Camilla Kazimi, and Dirk Van Amelsfort. 2003. "Drivers Willingness-To-Pay to Reduce Travel Time: Evidence from the San Diego I-15 Congestion Pricing Project." *Transportation Research Part A: Policy and Practice*, 37(4), 373 - 387.

Brownstone, David, and Thomas F. Golob. 1992. "The Effectiveness of Ridesharing Incentives: Discrete-Choice Models of Commuting in Southern California." *Regional Science and Urban Economics*, 22(1), 5 - 24.

Buchmann, Claudia, Thomas A. DiPrete, and Anne McDaniel. 2008. "Gender Inequalities in Education." *Annual Review of Sociology*, 34, 319 - 337.

Buliung, Ron, Randy Bui, and Ryan Lanyon. 2012. "When the Internet is Not Enough: Toward an Understanding of Carpool Services for Service Workers." *Transportation*, 39(5), 877 - 893.

Buliung, Ron, Kalina Soltys, Catherine Habel, and Ryan Lanyon. 2009. "Driving Factors Behind Successful Carpool Formation and Use." *Transportation Research Record: Journal of the Transportation Research Board*, 2118, 31 - 38.

Cai, Jing, Alain De Janvry, and Elisabeth Sadoulet. 2015. "Social Networks and the Decision to Insure." *American Economic Journal: Applied Economics*, 7(2), 81 - 108.

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou. 2009. "Peer Effects and Social Networks in Education." *Review of Economic Studies*, 76(4), 1239 - 1267.

Calvó-Armengol, Antoni, and Yves Zenou. 2009. "Social Networks and Crime Decisions: The Role of Social Structure in Facilitating Delinquent Behavior."

International Economic Review, 45(3), 939 - 958.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review*, 43(2), 245 - 257.

Carlsson, Magnus, and Dan-Olof Rooth. 2007. "Evidence of Ethnic Discrimination in the Swedish Labor Market using Experimental Data." *Labour Economics*, 14(4), 716 - 729.

Centola, Damon. 2011. "An Experimental Study of Homophily in the Adoption of Health Behavior." *Science*, 334(6060), 1269 - 1272.

Charles, Kerwin Kofi, and Patrick Kline. 2006. "Relational Costs and the Production of Social Capital: Evidence from Carpooling." *Economic Journal*, 116(511), 581 - 604.

Chin, Aimie, N. Meltem Daysal, and Scott A. Imberman. 2007. "Impact of Bilingual Education Programs on Limited English Proficient Students and Their Peers: Regression Discontinuity Evidence from Texas." *Journal of Public Economics*, 107, 63 - 78.

Chiswick, Barry R. 1978. "The Effect of Americanization on the Earnings of Foreign-Born Men." *Journal of Political Economy*, 86(5), 897 - 921.

Chiswick, Barry R. 1980. "The Earnings of White and Coloured Male Immigrants in Britain." *Economica*, 47(185), 81 - 87.

Chiswick, Barry R., and Paul W. Miller. 1994. "The Determinants of Post-Immigration Investments in Education." *Economics of Education Review*, 13(2), 163 - 177.

Chiswick, Barry R., and Paul W. Miller. 1995. "The Endogeneity between Language and Earnings: International Analyses." *Journal of Labor Economics*, 13(2), 246 - 288.

Chiswick, Barry R., and Paul W. Miller. 2001. "A Model of Destination-Language Acquisition: Application to Male Immigrants in Canada." *Demography*, 38(3), 391 - 409.

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine*, 357, 370 - 379.

Cline, Michael, Corey Sparks, and Karl Eschbach. 2009. "Understanding Carpool Use by Hispanics in Texas." *Transportation Research Record: Journal of the Transportation Research Board*, 2118, 39 - 46.

Cohen, Jere M. 1977. "Sources of Peer Group Homogeneity." *Sociology of Education*, 50(4), 227 - 241.

Cohen-Cole, Ethan, and Jason M. Fletcher. 2008. "Is Obesity Contagious? Social Networks vs. Environmental Factors in the Obesity Epidemic." *Journal of Health Economics*, 27(5), 1382 - 1387.

Coleman, James S., et al. 1966. *Equality of Educational Opportunity*. Washington DC: U.S. Government Printing Office.

Conley, Timothy G., and Christopher R. Udry. 2010. "Learning about a New Technology: Pineapple in Ghana." *American Economic Review*, 100(1), 35 - 69.

Constant, Amelie F., and Douglas S. Massey. 2003. "Self-Selection, Earnings, and Out-Migration: A Longitudinal Study of Immigrants to Germany." *Journal of Population Economics*, 16(4), 631- 653.

Constant, Amelie F., and Klaus F. Zimmermann. 2012. "The Dynamics of Repeat Migration: A Markov Chain Analysis." *International Migration Review*, 46(2), 362 - 388.

Correia, Gonçalo, and José Manuel Viegas. 2011. "Carpooling and Carpool Clubs: Clarifying Concepts and Assessing Value Enhancement Possibilities through a Stated Preference Web Survey in Lisbon, Portugal." *Transportation Research Part A: Policy and Practice*, 45(2), 81 - 90.

Cox, K. R. 1969. "The Voting Decision in a Spatial Context" *Progress in Geography*, 1, 81 - 117.

Crane, Randall. 2007. "Is There a Quiet Revolution in Women's Travel? Revisiting the Gender Gap in Commuting." *Journal of the American Planning Association*, 73(3), 298 - 316.

Crowley, Terry, and Claire Bowern. 2010. *An Introduction to Historical Linguistics*. New York: Oxford University Press.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica*, 77(4), 1003 - 1045.

Cutler, David M., and Edward L. Glaeser. 1997. "Are Ghettos Good or Bad?" *Quarterly Journal of Economics*, 112(3), 827 - 872.

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 2008. "Is the Melting Pot Still Hot? Explaining the Resurgence of Immigrant Segregation." *Review of Economics and Statistics*, 90(3), 478 - 497.

Damm, Anna P. 2009. "Ethnic Enclaves and Immigrant Labor Market Outcomes: Quasi-Experimental Evidence." *Journal of Labor Economics*, 27(2), 281 - 314.

Damm, Anna P. 2014. "Neighborhood Quality and Labor Market Outcomes: Evidence from Quasi-Random Neighborhood Assignment of Immigrants." *Journal of Urban Economics*, 79, 139 - 166.

Depken II, Craig A. 2002. "Free Agency and the Concentration of Player Talent in Major League Baseball." *Journal of Sports Economics*, 3(4), 335 - 353.

Drydakis, Nick. 2012. "Ethnic Discrimination in the Greek Housing Market." *Journal of Population Economics*, 24(4), 1235 - 1255.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics*, 118(3), 815 - 842.

Duncan, Brian, and Stephen J. Trejo. 2011. "Intermarriage and the Intergenerational Transmission of Ethnic Identity and Human Capital for Mexican Americans." *Journal of Labor Economics*, 29(2), 195 - 227.

Dustmann, Christian, and Ian Preston. 2001. "Attitudes to Ethic Minorities, Ethnic Context and Location Decisions." *Economic Journal*, 111(470), 353 - 373.

Edin, Per-Anders, Peter Fredriksson and Olof Aslund. 2003. "Ethnic Enclaves and the Economic Success of Immigrants: Evidence from a Natural Experiment." *Quarterly Journal of Economics*, 118(1), 329 - 357.

Edo, Anthony. 2002. "The Impact of Immigration on Native Wages and Employment." *B.E. Journal of Economic Analysis & Policy*, 15(3), 1151 - 1196.

Edwards, Rachel. 2006. "What's in a Name? Chinese Learners and the Practice of Adopting English Names." *Language, Culture and Curriculum*, 19(1), 90 - 106.

Edwards, Rosalind, and Chamion Caballero. 2008. "What's in a Name? An Exploration of the Significance of Personal Naming of Mixed Children for Parents from Different Racial, Ethnic and Faith Backgrounds." *Sociological Review*, 56(1), 39 - 60.

Espenshade, Thomas J., and Haishan Fu. 1997. "An Analysis of English-Language Proficiency among U.S. Immigrants." *American Sociological Review*, 114(4), 1102 - 1128.

Espinosa, Kristin E., and Douglas S. Massey. 1997. "Determinants of English Proficiency among Mexican Migrants to the United States." *International Migration Review*, 31(1), 28 - 50.

Farley, Reynolds, and John Haaga. 2005. *The American People: Census 2000*. New York: Russell Sage Foundation.

Feliciano, Cynthia. 2005. "Does Selective Migration Matter? Explaining Ethnic Disparities in Educational Attainment among Immigrants Children." *International Migration Review*, 39(4), 841 - 871.

Ferguson, Erik. 1991. "Ridesharing, Firm Size, and Urban Form." *Journal of Planning Education and Research*, 10(2), 131 - 142.

Ferguson, Erik. 1997. "The Rise and Fall of the American Carpool: 1970 — 1990." *Transportation*, 24(4), 349 - 376.

Foster, Gigi. 2006. "It's Not Your Peers, and It's Not Your Friends: Some Progress toward Understanding the Educational Peer Effect Mechanism." *Journal of Public Economics*, 90(8-9), 1455 - 1475.

Frank, Lawrence, Mark Bradley, Sarah Kavage, James Chapman, and T. Keith Lawton. 2008. "Urban Form, Travel Time, and Cost Relationships with Tour Complexity and Mode Choice." *Transportation*, 35(1), 37 - 54.

Freeman, Richard B. 2006. "People Flows in Globalization." *Journal of Economic Perspectives*, 20(2), 145 - 170.

Freeman, Richard B., and Wei Huang. 2015. "Collaborating with People Like Me: Ethnic Coauthorship within the United States." *Journal of Labor Economics*, 33(1), 289 - 318.

Frick, Bernd. 2009. "Globalization and Factor Mobility: The Impact of the "Bosman-Ruling" on Player Migration in Professional Soccer." *Journal of Sports Economics*, 10(1), 88 - 106.

Friesen, Jane, and Brian Krauth. 2010. "Sorting, Peers and Achievement of Aboriginal Students in British Columbia." *Canadian Journal of Economics*, 43(4), 1273 - 1301.

Fryer, Roland G., and Stephen D. Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics*, 119(3), 767 - 805.

Gao Yihong, Cheng Ying, Zhao Yuan, and Zhou Yan. 2005. "Self-Identity Changes and English Learning among Chinese Undergraduates." *World Englishes*, 24(1), 39 - 51.

Garip, Filiz. 2012. "Discovering Diverse Mechanisms of Migration: The Mexico-U.S. Stream from 1970 to 2000." *Population and Development Review*, 383(3), 393 - 433.

Gaviria, Alejandro, and Steven Raphael. 2001. "School-Based Peer Effects and Juvenile Behaviors." *Review of Economics and Statistics*, 83(2), 257 - 268.

Gerhards, Jürgen, and Silke Hans. 2009. "From Hasan to Herbert: Name-Giving Patterns of Immigrant Parents between Acculturation and Ethnic Maintenance." *American Journal of Sociology*, 114(4), 1102 - 1128.

Girard, Yann, Florian Hett, and Daniel Schunk. 2015. "How Individual Characteristics Shape the Structure of Social Networks." *Journal of Economic Behavior & Organization*. 115, 197 - 216.

Giuliano, Genevieve, Douglas W. Levine, and Roger F. Teal. 1990. "Impact of High Occupancy Vehicle Lanes on Carpooling Behavior." *Transportation*, 17(2), 159 - 177.

Glaeser, Edward L., and José A. Scheinkman. 2001. "Non-Market Interactions." NBER Working Paper No. 8053.

Glaeser, Edward L., José A. Scheinkman, and Bruce Sacerdote. 2003. "The Social Multiplier." *Journal of the European Economic Association*, 1(2/3), 345 - 353.

Gordon, Milton M. 1964. *Assimilation in American Life: The Role of Race, Religion, and National Origins*. New York: Oxford University Press.

Granovetter, Mark S. 1973. "The Strength of Weak Ties." *American Journal of Sociology*, 78(6), 1360 - 1380.

Granovetter, Mark S. 1985. "Economic Action and Social Structure: The Problem of Embeddedness." *American Journal of Sociology*, 91(3), 481 - 510.

Green, David A. 1999. "Immigrant Occupational Attainment: Assimilation and Mobility over Time." *Journal of Labor Economics*, 17(1), 49 - 79.

Greenwood, Michael J., and John M. McDowell. 1986. "The Factor Market Consequences of U.S. Immigration." *Journal of Economic Literature*, 24(4), 1738 - 1772.

Gross, Dominique M., and Nicolas Schmitt. 2003. "The Role of Cultural Clustering in Attracting New Immigrants." *Journal of Regional Science*, 43(2), 295 - 318.

Guven, Cahit, and Asadul Islam. 2015. "Age at Migration, Language Proficiency, and Socioeconomic Outcomes: Evidence From Australia." *Demography*, 52(2), 513 - 542.

Hao, Lingxin. 2004. "Private Support and Public Assistance for Immigrant Families." *Journal of Marriage and Family*, 65(1), 36 - 51.

Hellerstein, Judith K., Mark J. Kutzbach, and David Neumark. 2014. "Do Labor Market Networks Have an Important Spatial Dimension?" *Journal of Urban*

Economics, 79, 39 - 58.

Holmstrom, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics*, 10(1), 74 - 91.

Horowitz, Abraham D., and Jagdish N. Sheth. 1976. "Ride Sharing To Work: An Attitudinal Analysis." *Transportation Research Record: Journal of the Transportation Research Board*, 637, 1 - 8.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Paper No. 7867.

Huang, Hai-Jun, Hai Yang, and Michael G. H. Bell. 2000. "The Models and Economics of Carpools." *Annals of Regional Science*, 34(1), 55 - 68.

Iceland, John, Daniel Weinberg, and Lauren Hughes. 2014. "The Residential Segregation of Detailed Hispanic and Asian Groups in the United States: 1980-2010." *Demographic Research*, 31, 593 - 624.

Jackson, Matthew O., and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory*, 71(1), 44 - 74.

Johnson, Jacqueline S., and Elissa L. Newport. 1989. "Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language." *Cognitive Psychology*, 21(1), 60 - 99.

Jun, Myung-Jin. 2012. "The Effects of Seouls New-Town Development on Suburbanization and Mobility: A Counterfactual Approach." *Environment and Planning A*, 44(9), 2171 - 2190.

Kanbur, Ravi, and Xiaobo Zhang. 1999. "Which Regional Inequality? The Evolution of RuralCUrban and InlandCCoastal Inequality in China from 1983 to 1995." *Journal of Comparative Economics*, 27(4), 686 - 701.

Kandel, Denise B. 1978. "Homophily, Selection, and Socialization in Adolescent Friendships." *American Journal of Sociology*, 84(2), 427 - 436.

Kao, Grace, and Marta Tienda. 1995. "Optimism and Achievement: The Educational Performance of Immigrant Youth." *Social Science Quarterly*, 76(1), 1 - 19.

Kirdar, Murat G. 2009. "Labor Market Outcomes, Savings Accumulation, and Return Migration." *Labour Economics*, 16(4), 418 - 428.

Kleven, Henrik J., Camille Landais, and Emmanuel Saez. 2013. "Taxation and International Migration of Superstars: Evidence from the European Football Market." *American Economic Review*, 103(5), 1892 - 1924.

Kominski, Robert. 1989. "How Good is 'How Well'? An Examination of the Census English-Speaking Ability Question." Paper presented at the American Statistical Association Annual Meeting 1989, Washington DC, August 6 - 11.

Kopkin, Nolan. 2012. "Tax Avoidance: How Income Tax Rates Affect the Labor Migration Decisions of NBA Free Agents." *Journal of Sports Economics*, 13(6), 571 - 602.

Kremer, Michael, and Dan Lavy. 2008. "Peer Effects and Alcohol Use among College Students." *Journal of Economic Perspectives*, 22(3), 189 - 206.

Larkey, Linda Kathryn, Michael L. Hecht, and Judith Martin. 1993. "What's in a Name? African American Ethnic Identity Terms and Self-Determination." *Journal of Language and Social Psychology*, 12(4), 302 - 317.

Lazear, Edward P. 1995. "Culture and Language." NBER Working Paper No. 5249.

Lazear, Edward P. 2007. "Mexican Assimilation in the United States." In *Mexican Immigration to the United States*, eds., George J. Borjas. Chicago: University of Chicago Press.

Lee, Li Way. 1984. "The Economics of Carpooling." *Economic Inquiry*, 22(1), 128 - 135.

Lee, Wai-Sum, and Zee, Eric. 2003. "Standard Chinese (Beijing)." *Journal of the International Phonetic Association*, 33(1), 109 - 112.

Lenneberg, Eric H. 1967. *Biological Foundations of Language*. New York: Wiley.

Levels, Mark, Jaap Dronkers, and Gerbert Kraaykamp. 2008. "Immigrant Children's Educational Achievement in Western Countries: Origin, Destination, and Community Effects on Mathematical Performance." *American Sociological Review*, 73(5), 835 - 853.

Levinson, David M., and Ajay Kumar. 1994. "The Rational Locator: Why Travel Times Have Remained Stable." *Journal of the American Planning Association*, 60(3), 319 - 332.

Lichter, Daniel T. 2013. "Integration or Fragmentation? Racial Diversity and the American Future." *Demography*, 50, 359 - 391.

Lin, Xu, and Bruce A. Weinberg. 2014. "Unrequited Friendship? How Reciprocity Mediates Adolescent Peer Effects." *Regional Science and Urban Economics*, 48, 144 - 153.

Liu, Cathy Yang, and Gary Painter. 2012. "Travel Behavior among Latino Immigrants: The Role of Ethnic Concentration and Ethnic Employment." *Journal of Planning Education and Research*, 32(1), 62 - 80.

Lubotsky, Darren. 2007. "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings." *Journal of Political Economy*, 115(5), 820 - 867.

Lymperopoulou, Kitty. 2013. "The Area Determinants of the Location Choices of New Immigrants in England." *Environment and Planning A*, 45(3), 575 - 592.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60(3), 531 - 542.

Manski, Charles F. 2000. "Economic Analysis of Social Interactions." *Journal of Economic Perspectives*, 14(3), 115 - 136.

Marmaros, David, and Bruce Sacerdote. 2002. "Peer and Social Networks in Job Search." *European Economic Review*, 46(4-5), 870 - 879.

Marmaros, David, and Bruce Sacerdote. 2006. "How Do Friendships Form?" *Quarterly Journal of Economics*, 121(1), 79 - 119.

Massey, Douglas S., and Nancy A. Denton. 1985. "Spatial Assimilation as a Socioeconomic Outcome." *American Sociological Review*, 50(1), 94 - 106.

Massey, Douglas S., Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor. 1993. "Theories of International Migration: A

Review and Appraisal." *Population and Development Review*, 19(3), 431 - 466.

Mayer, Adalbert, and Steven L. Puller. 2008. "The Old Boy (and Girl) Network: Social Network Formation on University Campuses." *Journal of Public Economics*, 92(1-2), 329 - 347.

McManus, Walter, William Gould, and Finis Welch. 1983. "Earnings of Hispanic Men: The Role of English Language Proficiency." *Journal of Labor Economics*, 1(2), 101 - 130.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, 27, 415 - 444.

Meng, Xin, and Robert G. Gregory. 2005. "Intermarriage and the Economic Assimilation of Immigrants." *Journal of Labor Economics*, 23(1), 135 - 174.

Mizruchi, Mark S., and Linda Brewster Stearns. 2001. "Getting Deals Done: The Use of Social Networks in Bank Decision-Making." *American Sociological Review*, 66(5), 647 - 671.

Mok, Pegg P.K., and Volker Dellwo. 2008. "Comparing Native and Non-Native Speech Rhythm Using Acoustic Rhythmic Measures: Cantonese, Beijing Mandarin and English." *Proceedings of Speech Prosody, 2008*.

Montgomery, James D. 1991. "Social Networks and Labor-Market Outcomes: Toward an Economic Analysis." *American Economic Review*, 81(5), 1408 - 1418.

Mora, Toni, and Philip Oreopoulous. 2011. "Peer Effects on High School Aspirations: Evidence from a Sample of Close and Not-So-Close Friends." *Economics*

of Education Review, 30(4), 575 - 581.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish Émigrés and US Invention." *American Economic Review*, 104(10), 3222 - 3255.

Mosetti, Sauro, and Carmine Porello. 2010. "How Does Immigration Affect Native Internal Mobility? New Evidence from Italy." *Regional Science and Urban Economics*, 40(6), 427 - 439.

Mouw, Ted, and Yu Xie. 1999. "Bilingualism and the Academic Achievement of First- and Second-Generation Asian Americans: Accommodation with or without Assimilation?" *American Sociological Review*, 64(2), 232 - 252.

Munshi, Kaivan. 2003. "Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market." *Quarterly Journal of Economics*, 118(2), 549 - 599.

Munshi, Kaivan. 2004. "Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution." *Journal of Development Economics*, 73(1), 185 - 213.

Nicoll, Angus, Karen Bassett, and Stanley J. Ulijaszek. 1986. "What's in a Name? Accuracy of Using Surnames and Forenames in Ascribing Asian Ethnic Identity in English Populations." *Journal of Epidemiology and Community Health*, 40(4), 364 - 368.

Nyerges, Timothy, and Robert W. Aguirre. 2011. "Public Participation in Analytic-Deliberative Decision Making: Evaluating a Large-Group Online Field Experiment." *Annals of the Association of American Geographers*, 101(3), 561 - 586.

Oreopoulos, Philip. 2011. "Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes." *American Economic Journal: Economic Policy*, 3(4), 148 - 171.

Ottaviano, Gianmarco I.P. 2004. "Rethinking the Effect of Immigration on Wages." *Journal of the European Economic Association*, 10(1), 152 - 197.

Ottaviano, Gianmarco I.P., and Giovanni Peri. 2005. "Cities and Cultures." *Journal of Urban Economics*, 58(2), 304 - 337.

Ottaviano, Gianmarco I.P., and Giovanni Peri. 2006. "The Economic Value of Cultural Diversity: Evidence from US Cities." *Journal of Economic Geography*, 6(1), 9 - 44.

Patacchini, Eleonora, and Yves Zenou. 2012. "Urban Crime and Ethnicity." *Review of Network Economics*, 11(3).

Peri, Giovanni. 2012. "The Effect Of Immigration On Productivity: Evidence From U.S. States." *Review of Economics and Statistics*, 94(1), 348 - 358.

Peri, Giovanni, and Vasil Yasenov. 2015. "The Labor Market Effects of a Refugee Wave: Applying the Synthetic Control Method to the Mariel Boatlift." NBER Working Paper No. 21801.

Polavieja, Javier G. 2015. "Capturing Culture: A New Method to Estimate Exogenous Cultural Effects Using Migrant Populations." *American Sociological Review*, 80(1), 166 - 191.

Portes, Alejandro, and Min Zhou. 1993. "The New Second Generation: Segmented Assimilation and Its Variants." *American Academy of Political and Social*

Science, 530, 74 - 96.

Preston, Valerie, S. McLafferty, and X. F. Liu. 1998. "Geographical Barriers to Employment for American-born and Immigrant Workers." *Urban Studies*, 35(3), 529 - 545.

Qin, Yu, Hongjia Zhu, and Rong Zhu. 2016. "Changes in the Distribution of Land Prices in Urban China during 2007-2012." *Regional Science and Urban Economics*, 57, 77 - 90.

Rubinstein, Yona, and Dror Brenner. 2014. "Pride and Prejudice: Using Ethnic-Sounding Names and Inter-Ethnic Marriages to Identify Labour Market Discrimination." *Review of Economic Studies*, 81(1), 389 - 425.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2010. *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, 116(2), 681 - 704.

Sanchez, Thomas, Rich Stolz, and Jacinta Ma. 2004. "Inequitable Effects of Transportation Policies on Minorities." *Transportation Research Record: Journal of the Transportation Research Board*, 1885, 104 - 110.

Sanders, Jimy, Victor Nee, and Scott Sernau. 2002. "Asian Immigrants' Reliance on Social Ties in a Multiethnic Labor Market." *Social Forces*, 81(1), 281 - 314.

Schelling, Thomas C. 1969. "Models of Segregation." *American Economic Review*, 59(2), 488 - 493.

Scott, Darren M., Paul A. Coomes, and Alexei I. Izyumov. 2005. "The Location Choice of Employment-based Immigrants among U.S. Metro Areas." *Journal of Regional Science*, 45(1), 113 - 145.

Shaefer, David R., and Sandra D. Simpkins. 2014. "Using Social Network Analysis to Clarify the Role of Obesity in Selection of Adolescent Friends." *American Journal of Public Health*, 104(7), 1223 - 1229.

Shannon, Jerry. 2016. "Beyond the Supermarket Solution: Linking Food Deserts, Neighborhood Context, and Everyday Mobility." *Annals of the Association of American Geographers*, 106(1), 186 - 202.

Shifman, Limor, and Elihu Katz. 2005. "'Just Call Me Adonai': A Case Study of Ethnic Humor and Immigrant Assimilation." *American Sociological Review*, 70(5), 843 - 859.

Sinha, Paramita, and Maureen L. Cropper. 2013. "The Value of Climate Amenities: Evidence from US Migration Decisions." NBER Working Paper No. 18756.

Smart, Michael. 2010. "US Immigrants and Bicycling: Two-Wheeled in Autopia." *Transport Policy*, 17(3), 153 - 159.

Sobel, Michael E. 2000. "Causal Inference in the Social Sciences." *Journal of the American Statistical Association*, 95(450), 647 - 651.

Spence, Michael. 2015. "Job Market Signaling." *Quarterly Journal of Economics*, 87(3), 355 - 374.

Stark, Oded. 1991. *The Migration of Labor*, Oxford: Blackwell.

Stiefel, Leanna, Amy E. Schwartz, and Dylan Conger. 2004. "Age of Entry and the High School Performance of Immigrant Youth." *Journal of Urban Economics*, 67: 303 - 314.

Stinebrickner, Ralph, and Todd R. Stinebrickner. 2006. "What Can Be Learned About Peer Effects Using College Roommates? Evidence From New Survey Data and Students from Disadvantaged Backgrounds." *Journal of Public Economics*, 90(8-9) 1435 - 1454.

Sue, Christina A., and Edward E. Telles. 2007. "Assimilation and Gender in Naming." *American Journal of Sociology*, 112(5), 1383 - 1415.

Tainer, Evelina. 1988. "English Language Proficiency and the Determination of Earnings among Foreign-Born Men." *Journal of Human Resources*, 23(1), 108 - 122.

Teal, Roger F. 1987. "Carpooling: Who, How and Why." *Transportation Research Part A: General*, 21(3), 203 - 214.

Tella, Adeyinka. 2014. *Social Media Strategies for Dynamic Library Service Development*. Hershey: IGI Global.

Trogdon, Justin G., James Nonnemaker, and Joanne Pais. 2008. "Peer Effects in Adolescent Overweight." *Journal of Health Economics*, 27(5), 1388 - 1399.

Topa, Giorgio. 2001. "Social Interactions, Local Spillovers and Unemployment." *Review of Economic Studies*, 68(2), 261 - 295.

Urquiola, Miguel. 2005. "Does School Choice Lead to Sorting? Evidence from Tiebout Variation." *American Economic Review*, 95(4), 1310 - 1326.

Verdier, Thierry, and Yves Zenou. 2015. "The Role of Social Networks in Cultural Assimilation." IZA Discussion Paper No. 9341.

Viton, Philip A. 1992. "On Frontier Specifications for Discrete Binary Choice Analysis." *Journal of Regional Science*, 32(3), 285 - 308.

Wimmer, Andreas, and Kevin Lewis. 2010. "Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook." *American Journal of Sociology*, 116(2), 583 - 642.

Yang, Hai, and Hai-Jun Huang. 1999. "Carpooling and Congestion Pricing in a Multilane Highway with High-Occupancy-Vehicle Lanes." *Transportation Research Part A: Policy and Practice*, 33(2), 139 - 155.

Zee, Eric. 1985. "Sound Change in Syllable Final Nasal Consonants in Chinese." *Journal of Chinese Linguistics*, 13(2), 291 - 330.

Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *Review of Economics and Statistics*, 85(1), 9 - 23

Zhou, Min. 1997. "Segmented Assimilation: Issues, Controversies, and Recent Research on the New Second Generation." *International Migration Review*, 31(4), 975 - 1008.

Zhou, Min, and Carl L. Bankston III. 1994. "Social Capital and the Adaptation of the Second Generation: The Case of Vietnamese Youth in New Orleans." *International Migration Review*, 28(4), 821 - 845.

Zussman, Asaf. 2013. "Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars." *Economic Journal*, 123(572), F433 - F468.